Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

Outi Pönni

Post-processing wind speed forecasts with the extended logistic regression method for energy production

Master's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology in the Degree Programme in Engineering Physics and Mathematics.
Espoo, October 20, 2014

Supervisor: Professor Ahti Salo, Aalto University
Advisor: Ph. Lic. Juha Kilpinen

Aalto University
School of Science
Degree Programme in Engineering Physics and Mathematics

ABSTRACT OF
MASTER'S THESIS

Author: Outi Pönni

Title:
Post-processing wind speed forecasts with the extended logistic regression method for energy production

Date:	October 20, 2014	**Pages:**	vii + 73
Major:	Systems and Operations Research	**Code:**	Mat-2
Supervisor:	Professor Ahti Salo		
Advisor:	Ph. Lic. Juha Kilpinen		

Wind speed forecasts are used to forecast wind electricity production. More accurate wind speed forecasts reduce uncertainty in electricity production forecasts and therefore also balancing costs. In this thesis, we investigate the extended logistic regression (ELR) method for forecasting wind speeds by using average numerical AROME/Harmonie forecasts as the initial forecasts. The ELR produces a probability distribution, which is converted into wind speed forecasts. Ideally, the ELR produces wind speed forecasts with less uncertainty.

The ELR model is created for a wind mast situated in Olkiluoto in southwestern Finland. The short-term (3h) forecasts are made at 03 UTC each day and the data spans eight months. Most of the data is used as the training data while a period of one month marks the verification data. The best ELR model is chosen among alternative ELR models involving different parameter combinations according to verification scores and statistical results. In this thesis, we investigate whether the ELR method produces more accurate forecasts with less uncertainty in this model composition than the initial forecast or the reference models.

The best ELR model includes the square root of the threshold wind speed, wind speed and wind direction as parameters. The wind speed forecasts created with this ELR model reduce the absolute mean error by 29% and the root mean square error by 24% compared to the AROME/Harmonie forecasts. Furthermore, the graphical analysis supports the choice of this ELR model. Consequently, the ELR method is adequate for wind speed forecasting and it can be developed further. The results can be used for commercial use after the model is extended to cover all hours of a day and for longer AROME/Harmonie forecasts. Models can also be created for other locations with observation data available. In the future, the model can be extended to cover all Finland.

Keywords: Extended logistic regression, probabilistic forecast, wind speed forecast, verification

Language: English

Aalto-yliopisto
Perustieteiden korkeakoulu
Teknillisen fysiikan ja matematiikan koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Outi Pönni		
Työn nimi:	Tuuliennusteiden jälkiprosessointi laajennetulla logistisella regressiomallilla energiantuotantoa varten		
Päiväys:	20. lokakuuta 2014	**Sivumäärä:**	vii + 73
Pääaine:	Systeemi- ja operaatiotutkimus	**Koodi:**	Mat-2
Valvoja:	Professori Ahti Salo		
Ohjaaja:	FL Juha Kilpinen		

Tuulen nopeusennusteita tarvitaan tuulienergian tuotannon ennustamiseen. Tarkemmat tuulen nopeusennusteet vähentävät epävarmuutta tuulienergiaennusteissa ja vähentävät siten sähkön säätökustannuksia. Tässä diplomityössä tutkimme laajennettua logistista regressiomenetelmää (ELR) tuulen nopeuden ennustamiseen käyttäen alkuperäisenä ennusteena keskimääräistä numeerista AROME/Harmonie -ennustetta. ELR tuottaa todennäköisyysjakauman, joka muutetaan tuulen nopeusennusteeksi.

ELR-malli luodaan Olkiluodossa sijaitsevalle tuulimastolle käyttäen kahdeksan kuukauden pituista datajaksoa. Malli ennustaa lyhytaikaista (3h) tuulen nopeutta klo 03 UTC joka vuorokausi. ELR-malli muodostetaan käyttäen harjoitusdatana suurinta osaa datasta ja verifiointidatana yhden kuukauden dataa. Paras ELR-malli valitaan vaihtoehtoisista eri parametrikombinaatioita sisältävistä malleista verifiointiarvojen ja tilastollisten suureiden perusteella. Diplomityössä tutkitaan, tuottaako ELR-metodi tarkempia ennusteita kuin alkuperäinen ennuste tai referenssimallit tässä malliasetelmassa.

Paras ELR-malli käyttää parametreina tuulen kynnysnopeuden neliöjuurta, tuulen nopeutta sekä tuulen suuntaa. ELR-mallilla saatujen tuulen nopeusennusteiden absoluuttinen keskivirhe on 29% pienempi ja neliöllinen virhe 24% pienempi kuin AROME/Harmonie ennusteiden. Myös graafinen tarkastelu tukee ELR-mallin valintaa. ELR-metodi toimii siis hyvin tuulen nopeuden ennustamiseen ja ELR-metodia kannattaa tutkia laajemmin. Tuloksia voi hyödyntää kaupallisesti sitten, kun malli on laajennettu käsittämään kaikki vuorokauden tunnit ja pidempiä ennusteita on tehty AROME/Harmonie -dataa käyttäen. Malleja voi myös luoda muille tuulimastoille, joille on käytössä havaintodataa. Tulevaisuudessa malli voidaan laajentaa kattamaan koko Suomi.

Asiasanat:	Laajennettu logistinen regressio, todennäköisyysennuste, tuuliennuste, verifiointi
Kieli:	Englanti

Acknowledgements

First of all, I thank my instructor Ph. Lic. Juha Kilpinen from the Finnish Meteorological Institute for the interesting and motivating topic and all the guidance during my thesis. It was a privilege to work with Juha and he made it possible to have such good results thanks to all the background work of Juha including investigation of potential methods, research articles to get started with the topic and the preliminary code for R that was the basis for the model. Juha gave me precious advice, which directed me to make choices in the thesis and practical problems with understanding the data, the method and the R code. Shortly, I am grateful I had the opportunity to work with Juha and make my thesis in this field.

Secondly, I thank my Professor Ahti Salo from Aalto University Systems Analysis Laboratory for the significant help with writing the thesis involving essential comments as well as critical inspection of the choices made in the thesis. In addition to guidance with scientific writing and relevant references, Ahti advised me with the perspective of the thesis. Ahti gave me valuable support during the thesis, which enabled text and analysis of high quality.

Thirdly, I thank my colleagues at the Finnish Meteorological Institute, especially Evgeny Atlaskin for the data, Irene Suomi for the help with the stability parameter, Jenni Latikka for suggesting me articles about wind energy, as well as Pertti Nurmi and Markku Kangas for borrowing me books of verification. Moreover, I thank Zana Cranmer and Hannele Holttinen for sharing precious knowledge of wind power.

In addition, I thank Ulla Aarnio and Max Andersson for pre-reading the thesis and Heikki Pönni and Marja Aarnio-Frisk for valuable advice. They all gave me support and encouragement during this laborious project.

Espoo, October 20, 2014

Outi Pönni

Acronyms

ALADIN	Aire limitée adaptation dynamique initialisation
AROME	Application of research to operations at mesoscale
BS	Brier score
BSS	Brier skill score
CET	Central European time
ECMWF	European centre for medium-range weather forecasts
ELR	Extended logistic regression
EU	European union
HIRLAM	High resolution limited area model
LAM	Limited area model
LM	Linear model
LR	Logistic regression
MAE	Mean absolute error
ME	Mean error
MOS	Model output statistics
MSE	Mean square error
NWP	Numerical weather prediction
REEP	Regression estimation of event probabilities
RMSE	Root mean square error
ROC	Relative operating characteristics
STDE	Standard deviation
TSO	Transmission system operator
UTC	Coordinated universal time

Contents

Chapter 1

Introduction

1.1 Problem statement

Wind speed is variable and unpredictable and therefore difficult to forecast accurately. The atmospherical equations describing wind speed are complex and incomplete and there is no perfect model to describe the behaviour of wind. The purpose of this thesis is to create a model to forecast wind speed better and with less uncertainty than the existing model used in the Finnish Meteorological Institute.

A method called extended logistic regression (ELR) will be used and compared to two simple methods; regression estimation of event probabilities (REEP) and linear model (LM). The ELR is a nonlinear regression method that gives a probability distribution of wind speed at each measurement time. The ELR model will be created by post-processing numerical wind speed forecasts calculated with supercomputers. These numerical forecasts are average forecasts for a certain area and thus not very accurate forecasts for specific points. In this thesis, we will model the wind speed of a specific wind mast so that the numerical forecasts will be post-processed to match better with the observations of the wind mast.

To improve reliability, the numerical forecast and observation data should be gathered from as long a period as possible. In this thesis, the available data spans eight months. Most of the data will be used for creating the model and the rest of the data will be used for testing the model. Verification scores describe the difference between the probabilistic forecasts and the observations whereas basic error scores describe the difference between the actual wind speed forecasts and the observed wind speeds. Verification scores for the alternative models are compared to determine if the chosen model works better than other models and the basic error scores are analysed to compare the new forecasts and the initial forecasts.

In this thesis, we focus on wind speed forecasting in Finland and use the data from the Finnish Meteorological Institute. The ELR model will be developed for a wind mast situated in Olkiluoto, a coastal area in southwestern Finland. The model height of 60m is chosen since new large commercial wind turbines are tall and this height is therefore good for power production. The initial forecast length ranges from 3 to 24 hours. The effects of the model parameters are more noticeable with a short forecast length so the forecast length of 3 hours is chosen to make it is easier to validate the results. However, the models for other heights and different forecast lengths would be very similar.

Wind speed forecasts are needed to forecast wind power. The amount of wind power depends on the cube of wind speed so the production varies according to wind speed. The share of wind power in electricity production is increasing making it important to have accurate wind power forecasts. Consequently, the created forecasts will be investigated from the point of view of wind power production.

1.2 Scope and structure

The ELR model created will be very specific and only forecast wind speed for the Olkiluoto wind mast at the height of 60m above the ground level. The data available spans approximately eight months from 1st January 2010 to 17th August 2010. The ELR model will forecast short-term (3h) wind speed for each day at 03 UTC and give a probability distribution, which will be converted into wind speed values. Due to limited amount of data, the model will intentionally not be too complex so that parameters can be physically justified and the impact of each parameter will be significant. The purpose is to investigate if the ELR model will improve the initial numerical forecasts and be superior to the reference models. Both the functionality of the ELR model created and the use of the ELR method for forecasting wind speeds will be analysed.

In this thesis, we cover the background, creating and verifying the model and analysing the results. Chapter 2 introduces the trend of renewable energies in Finland, basics about wind power and electricity markets, and the significance of accurate wind speed forecast in wind power production. In Chapter 3, we explain weather models and numerical forecasts and also introduce the data. Chapter 4 presents the theory of the extended logistic regression method and the reference methods, as well as verification and statistical methods for the testing of the results. Chapter 5 involves describing the implementation and the results of the ELR model and the reference models, and the chosen model is then evaluated in Chapter 6. The conclusions are made together with discussing the results in the light of power production in Chapter 7.

Chapter 2

Wind power

2.1 Renewable energies

The amount of renewable energy production is increasing in the world. This is due to both technological development, climate policies, increasing energy consumption and need for self-sufficiency in energy production. New technologies are invented to make energy production more efficient and new sites for energy production are investigated. Along with the growth of renewable energies, the amount of wind power produced is growing. Wind power is becoming more important in the energy production because wind is an inexhaustible natural resource with low maintenance costs and it can be produced in various locations. Wind power forecasts play a significant role when electricity generated with wind power is sold into the grid.

In 2013, the electricity production in Finland was 68.2TWh with the share of renewable energy being 36% (Finnish Energy Industries, 2014). The relative proportions of electricity production types in Finland are exemplified in Figure 2.1. The share of wind power was only 1.1% in 2013 being much less than the share of water power (18.7%) and biomass (15.7%) but bigger than the share of solar power that was too small to be taken into calculations. Nevertheless, Finnish Energy Industries (2014) states that the share of wind power will increase, estimating the share of wind power being $10 - 15\%$ in 2050. It means around $15 - 20$TWh/year produced with wind power. A reason for the forthcoming increase of the share of wind power is the environmental value of wind power. Wind power produces no emissions, which makes it preferred to fossil fuel based energy sources (Holttinen et al., 2011). The greenhouse effect and the climate change encourage environmentally conscious behaviour and the choice of wind power. With no fuel costs wind power also reduces the total operating costs of the power system (Holttinen et al., 2011), which makes it an economical alternative

for electricity markets.

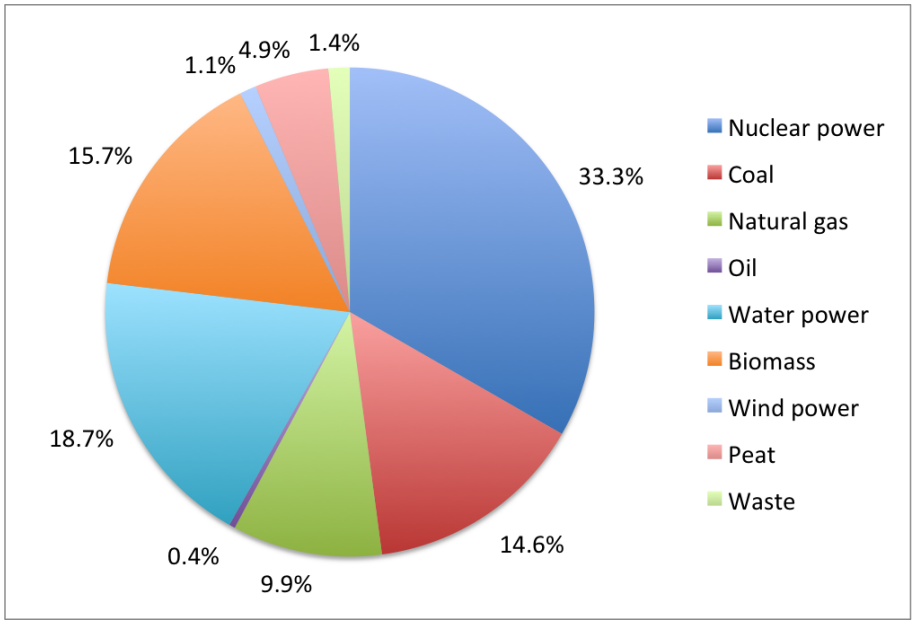


Figure 2.1: The shares of the different electricity production types in Finland in 2013 (Finnish Energy Industries, 2014).

Policies have a major effect on new energy investments. The latest national energy and climate targets for Finland set in the Finnish National Energy and Climate Strategy interconnect with the targets of the European Union (EU). The EU obligates Finland to produce 38% of energy with renewable energy sources by 2020 (The Ministry of Employment and the Economy (Finland), 2013a). Finland is reaching this objective and already now Finland exceeds the minimum amount of renewable energies set by the EU. Additionally, the strategy defines that the construction of wind power plants in Finland should be sped up with the aim of 9TWh by 2025. The previously set target is to produce 6TWh with wind power by 2020 (The Ministry of Employment and the Economy (Finland), 2013a).

To conclude, The Ministry of Employment and the Economy (Finland) (2013a) suggests that the amount of renewable energies is going to grow in Finland. Furthermore, Finland tries to facilitate the construction of wind power to raise the amount of wind power plants. Thus, wind power production will be more important in the near future. The increase of wind power makes it even more significant to forecast wind power correctly following the increased amount of wind power inserted to the grid.

2.2 Characteristics of wind power

Wind is a renewable source of energy. Wind energy is created by converting the kinetic energy of air flows first into rotation and then into electricity with wind turbines and generators that are connected to an electrical network. Wind power is directly proportional to the cube of wind speed according to the equation

$$P = \frac{\rho}{2} A u^3,$$

where P is the power generated, ρ is the air density, A is the perpendicular area to the wind and u is the wind speed (Manwell et al., 2009). However, rotors of a wind turbine can not exploit all the wind power. The maximum theoretically usable wind power (Manwell et al., 2009) is

$$P = \frac{16}{27} \frac{\rho}{2} A u^3.$$

Wind energy is therefore

$$E = Pt = \frac{16}{27} \frac{\rho}{2} A u^3 t,$$

where t is time. Wind power curves as a function of wind speed are illustrated in Figure 2.2 (The Ministry of Employment and the Economy (Finland), 2013b)[1]. The real wind power curves differ from the maximum theoretically usable wind power curve due to disturbances, such as ice on the rotors (The Ministry of Employment and the Economy (Finland), 2013b) or wake effects (Manwell et al., 2009). The real power curves in Figure 2.2 display that a wind turbine connected to the grid starts at about 4m/s, called cut-in wind speed, and shuts down at about 25m/s, called cut-off wind speed (The Finnish Wind Power Association, 2014). Big turbines reach the maximum power, also called nominal or rated power, at about 10 – 15m/s. Wind power is not produced with wind speeds greater than 25m/s for safety reasons. Therefore, storms have to be forecasted to know when the production stops.

The production of wind power varies according to wind speed, which leads to challenges when inserting wind power to the electrical grid. Wind power can be stored with batteries, but that is not economical because it is cheaper to produce electricity from natural gas than to storage electricity with costly batteries (Busby, 2012). Therefore, wind power must be used immediately. It causes problems since wind power is not

[1] Texts in the figure are translated into English.

easy to predict accurately and so electricity production is insecure.

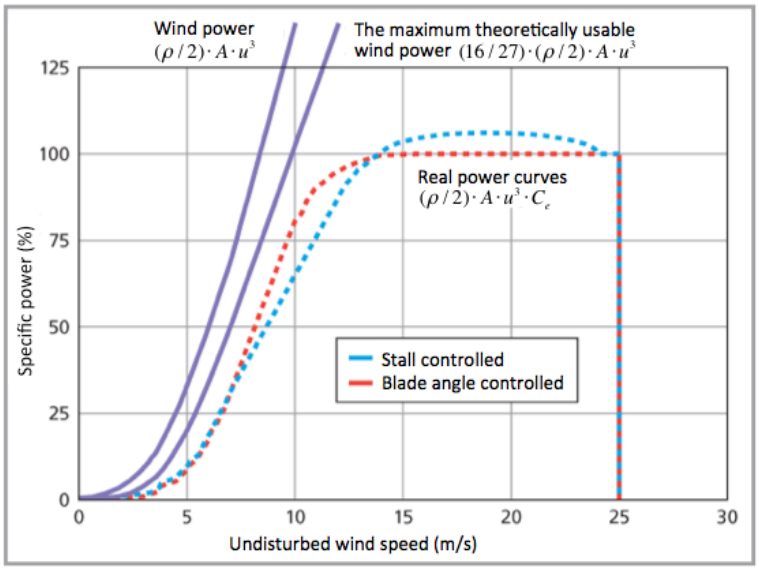


Figure 2.2: The wind power curve, the maximum theoretically usable wind power curve and power curves for turbines with fixed and adjustable blades. C_e stands for the efficiency of the generator.

Wind speed increases as a function of altitude and wind energy generated increases with the rotor diameter. Consequently, the size of the wind turbines is growing as showed in Figure 2.3 (The Ministry of Employment and the Economy (Finland), 2013b). Thus, we are interested in high economical wind turbines in this thesis.

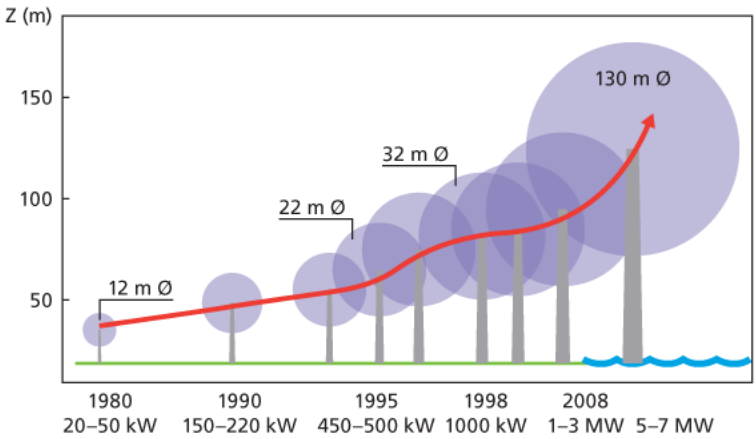


Figure 2.3: The trend of the wind turbine size.

2.3 Wind speed forecasts in power production

Wind power production is sold to the electricity markets. Varying with wind speed, the production can only be controlled by restricting it, which is not economically reasonable as long as there are no production costs, the spot prices are positive and subsidies depend on production (Holttinen et al., 2013). Since the wind power producer cannot decide the amount of power generated, production forecasts are needed. Wind energy is directly proportional to the cube of wind speed so the wind speed forecasts need to be as accurate as possible to minimise the regulation balancing costs resulting from forecast errors. Accurate forecasts benefit both the producer and the system operator.

Nord Pool Spot (Nord Pool Spot, 2014) operates the physical electricity trade in the Nordic and Baltic countries. Nord Pool Spot is divided into two markets: the day-ahead market Elspot and the intraday market Elbas. The bids for the Elspot market for the next day must be done latest at 12.00CET, which is 12 hours before the first delivery hour. Thus, the hourly wind power forecasts should be made 12 – 36 hours ahead. Each electricity producer makes a bid for how much electricity it wants to sell each hour and for what price. The final electricity price is then determined by supply and demand curves at Elspot after the market has closed. The intraday market Elbas is basically meant for incidents and changes in production. It is a continuous market where the bids are submitted one hour before the delivery hour.

The transmission system operator (TSO) (Nord Pool Spot, 2014) is responsible for the security of supply and operates the regulating power market to keep the transmission grid stable. If the amount of electricity generated is too low in relation to consumption, up-regulation is needed. In practice, the electricity in the grid is increased with balancing power the TSO buys from the producers that have extra generation capacity. Respectively, if the amount of electricity generated is too high in relation to consumption, down-regulation is needed and the TSO has to ensure that one or more producers reduce their electricity generation.

As long as there is a balance in Nord Pool Spot, the producer gets no penalty costs regardless of the electricity forecast (Holttinen et al., 2013). The producer, who predicts the amount of electricity generated incorrectly, still has to buy the missing electricity or sell the surplus electricity at the spot price. If the market is off-balance and the forecast is incorrect, the producer has to pay balancing costs. Balancing costs were approximately 3€/MWh less revenue for each produced MWh for one wind power site in 2011 and 2012 (Holttinen et al., 2013). Balancing costs for aggregated sites with shared prediction error were approximately 1 – 1.4€/MWh less revenue in 2011 and

2012, which is 60% smaller than the costs for individual sites (Holttinen et al., 2013). Although aggregation of sites decrease balancing costs, the costs can still be minimised by modelling wind speed better. Smaller prediction errors lead to less balancing and lower balancing costs.

Balancing costs vary according to the hour and the size of the error therefore being complicated to calculate the benefit of better forecasts. The power forecasts are not simple to compute from wind speed forecasts either because a wind turbine with a large rotor diameter catches several different wind speeds and the real production is only estimated. A rough estimate for the total cost for producers in Finland with 6TWh production is 0.6 – 0.84M€ less revenue, assuming aggregation of sites and 10% error in wind speed forecasts causing 10% error in wind production forecasts. Thus, it is economically wise to minimise errors of wind speed forecasts in a large scale.

The Elbas market could be a good choice for wind electricity trade because wind power is difficult to predict far ahead. However, it is not cost-effective to trade wind power in the Elbas market with low wind power penetration levels as long as the balancing costs are small since trading costs may outweigh profits (Holttinen and Koreneff, 2012). There are no extra regulating costs with wind power penetration being less than 10% of gross demand in Finland (Holttinen, 2004) and approximately 50% of the imbalances are ignored because the share of wind power in Finland is so small (Holttinen et al., 2013). With wind power penetration being higher than 10%, as in Denmark, the regulating costs are bigger and the Elbas market is cost-effective (Holttinen, 2004). Therefore, trading in the Elbas market will be economic for Finland in 2020 if the target to produce 6TWh with wind power is achieved. It means that the three hours' short-term forecasts created in this thesis will likely become useful in the near future. If it turns out that the extended logistic regression method improves the initial wind speed forecasts, the method can also be used for making longer forecasts useful for the Elspot market.

The grid load is essential from the system operator's point of view. The system operator has to manage the grid carefully and get accurate information about wind power production and electricity demand. Wind power affects the grid balance because of common prediction errors. Consequently, wind power increases the need for short-term reserve capacity (Holttinen et al., 2011). Wind power is not a problem as long as the share of wind power is low compared to other energy sources. Since the share of wind power is increasing, the need for regulating capacity will increase as well. More accurate forecasts reduce the need for regulating capacity so there is a call for developing forecast models.

Chapter 3

Forecasting wind speeds

3.1 Weather models

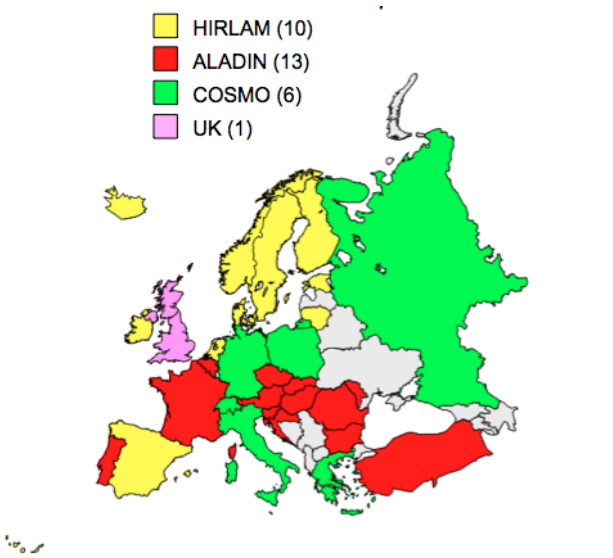
There are two types of atmospheric models: climate models and weather models. Climate models describe long-term (20 – 100 years) behaviour of atmosphere while weather models describe short-term (1 – 15 days) quick weather variations. In this thesis, we are interested in weather models and short-term weather prediction.

The weather models used in the Finnish Meteorological Institute are described in the Finnish Wind Atlas (The Ministry of Employment and the Economy (Finland), 2013b). Weather models compute weather characteristics, such as air pressure, wind, temperature, humidity, clouds, rain and sunshine, from complex physical equations to create a dynamical model that describes the atmosphere at each moment. Forecasts made with a weather model are not flawless because the model calculations are simplified and a forecast is made for a large area, not for a specific point. A weather model divides the atmosphere into a three dimensional grid and computes average conditions of the atmosphere for each area within grid points. The distance between grid points is called resolution.

Furthermore, weather models are categorised as global models and limited area models (LAM). Global models extend the entire atmosphere of the planet while LAMs cover a restricted domain of the atmosphere as illustrated in Figure 3.1(a). LAMs attain a better local accuracy with the same computer resources by focusing on a smaller area. Global models and LAMs are linked as global models create the boundaries of LAM models. That is, global models describe the behaviour of the weather at the borders of the LAMs. There are different LAMs made in cooperation between meteorological centres in Europe. The LAM consortiums in Europe in 2011 are presented in Figure 3.1(b). The Finnish Meteorological Institute is part of the High

resolution limited area model (HIRLAM) consortium. The other member countries of HIRLAM (HIRLAM, 2013) are Denmark, Estonia, Iceland, Ireland, the Netherlands, Norway, Spain, Sweden and Lithuania, France being an associate member.

(a) A global model versus a limited area model.



(b) LAM consortiums in Europe illustrated with different colours.

Figure 3.1: Limited area models. (The Finnish Meteorological Institute, 2014b)

3.2 Numerical weather prediction

Weather models are created by using numerical weather prediction (NWP) method (e.g. Coiffier, 2011). NWP is a numerical technique to solve the conditions of the atmosphere from complicated equations that describe fluid mechanics. Since computers cannot solve differential equations, the continuous equations are discretised, in other words, altered to a numerical form. Consequently, continuous functions are estimated with a discrete function with a finite number of values. NWP methods also include approximations and parameterisations. Parameterisations describe the characteristics of the atmosphere that are too small, too complex or not well understood to be expressed accurately.

All the NWP methods are based on the same continuous equations, called Euler equations, representing the behaviour of a non-viscous fluid. The equations are the

momentum equation, the thermodynamic equation, the continuity equation, the water vapour equation and the equation of state. The general meteorological approach to Euler equations is to form non-hydrostatic equations by a few simplifications, such as assessing the atmosphere as a thin layer. This approach is advantageous for modelling mesoscale progress of the atmosphere. Another approach is to form hydrostatic equations, also known as primitive equations, by setting vertical acceleration to zero. This approach is used extensively in weather forecasting for synoptic scale models. For a synoptic scale, the vertical velocities are much smaller than horizontal velocities, what allows neglecting vertical acceleration and enables the formation of a hydrostatic balance.

To create the NWP forecasts, the model is given initial values. However, the available observations can be spatially or temporally dispersed, which causes a problem with the initial data input in the model. This problem is resolved by data assimilation, illustrated in Figure 3.2. Data assimilation is a process in which observations and earlier forecasts, called first guess field, are combined to get a better initial state for the model. The further the data assimilation proceeds, the more accurate the initial state. Data assimilation can be performed for example in 3 or 6 hour cycles, where the first guess field is taken at the end of each cycle. Forecast errors often originate from the initial observation data and grow during calculations. The benefit of data assimilation is that it reduces the root mean square error of forecasts (Seity et al., 2010).

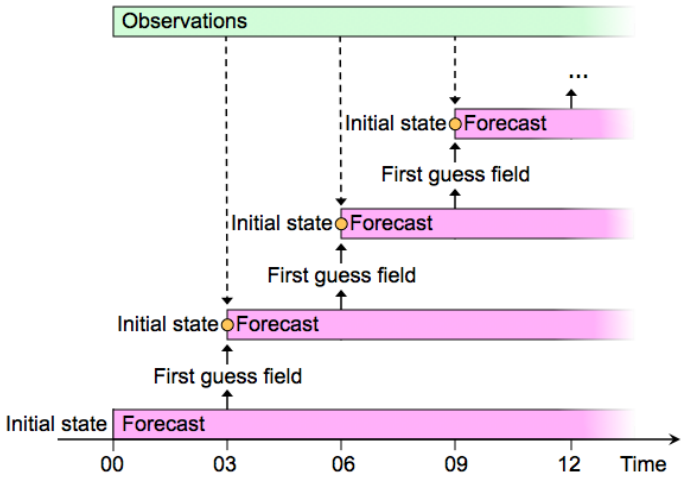


Figure 3.2: Data assimilation for 3h continuous assimilation cycle is exemplified with orange dots.

3.3 Limited area models

Limited area models used in the Finnish Meteorological Institute are the high resolution limited area model (HIRLAM), the application of research to operations at mesoscale (AROME) and an extension of the AROME called AROME/Harmonie. The HIRLAM (HIRLAM, 2013) is a hydrostatic synoptic scale weather model whereas the AROME/Har- monie used in this thesis is a mesoscale model. The synoptic scale refers to 5 – 15km horizontal resolution whereas the mesoscale refers to more accurate target resolutions. The European centre for medium-range weather forecasts (ECMWF) is a global model, which creates the boundaries for both HIRLAM, AROME and AROME/Harmonie (HIRLAM, 2013).

The latest version of the HIRLAM, version 7.4 (HIRLAM, 2013), has a horizontal resolution of 7.5km and 65 vertical levels with the lowest level at 12m. The HIRLAM uses the following prognostic variables: two horizontal wind components, vertical wind speed, temperature, specific humidity, geopotential height and pressure. The physical parameterisations of the HIRLAM are radiation, deep convection, turbulence, surface and microphysics. Data assimilation scheme used in the HIRLAM is a four-dimensional variational assimilation scheme with 6 hours assimilation cycle. The HIRLAM is run four times a day, at 00UTC, 06UTC, 12UTC and 18UTC and it is possible to make up to 54 hour forecasts (The Finnish Meteorological Institute, 2014a).

The AROME (Seity et al., 2010) is a LAM developed by Météo France in 2008. AROME is a non-hydrostatic version of the Aire Limitée Adaptation Dynamique Développement International (ALADIN), which is a LAM of Météo France covering western Europe. The AROME uses the ALADIN for the adiabatic part of the Euler equations. The AROME takes parameterisations for the physical parts of the Euler equations from the mesoscale non-hydrostatic Méso-NH model. The objective was to get best possible components available for the AROME (Seity et al., 2010). The Hirlam Aladin regional/meso-scale operational NWP in Europe (HARMONIE) was created in cooperation between the HIRLAM and the ALADIN in 2011. The HARMONIE (HIRLAM, 2013) is basically the AROME model that works for all member countries of the HIRLAM including Finland.

In this thesis, by referring to the AROME/Harmonie model, we mean the specific AROME/Harmonie model for Finland introduced in the Finnish Wind Atlas (The Ministry of Employment and the Economy (Finland), 2013b). The AROME/Harmonie covers the area presented in Figure 3.3 (The Finnish Meteorological Institute, 2014b) and describes the atmospheric characteristics of Finland and its immediate surroundings. The horizontal resolution of the AROME/Harmonie is relatively sharp, 2.5km,

with 300 grid points in the east–west direction and 600 grid points in north–south direction. There are 40 vertical levels with 30m as the lowest level. The model calculations are performed at minute intervals. The AROME/Harmonie uses a three-dimensional variational assimilation scheme to produce the initial state for the model (Seity et al., 2010). Data assimilation is implemented continuously in 3h cycles. The discretisation is performed with a two time level, semi-implicit semi-Lagrangian discretisation method.

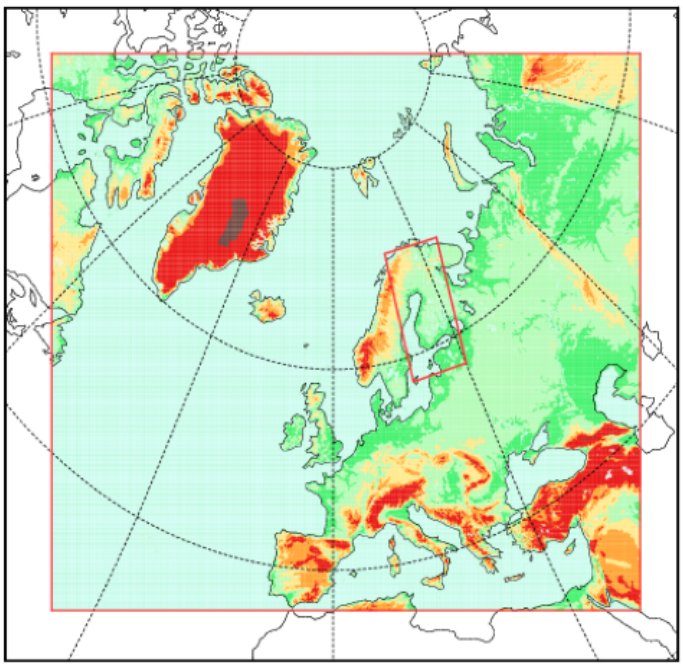


Figure 3.3: The red lines around Europe illustrate the boundaries of the HIRLAM and the red lines around Finland illustrate the boundaries of the AROME/Harmonie LAM.

The AROME/Harmonie calculates two components of horizontal wind speed, vertical wind speed, air pressure and temperature among other things. The quantities parameterised in the AROME/Harmonie model are radiation, interactions with the Earth's surface, boundary layer and turbulent diffusion, microphysics and shallow convection (Coiffier, 2011). The AROME/Harmonie separates water, normal ground and ground covered with ice, snow or vegetation and built areas (The Ministry of Employment and the Economy (Finland), 2013b). Altogether, there are four different types of surface — sea, lakes and rivers, nature and city — that are separated into subtypes. For example, nature has 12 subtypes from bogs to fields. Each type of surface is linked

with a quantity called roughness. The less obstacles for wind, the smaller the roughness. The value for each grid area is calculated as the weighted average according to the shares of the types.

3.4 Data

The AROME/Harmonie forecast data that is post-processed in this thesis is computed for the area where the Olkiluoto mast is located. The AROME/Harmonie data includes vertical wind speed, wind direction and temperature at several heights: 30m, 47m, 50m, 60m, 90m, 100m, 131m and 143m. Wind speed is expressed in meters per second (m/s), wind direction in degrees (°) and temperature in Kelvin (K). The accuracy of each magnitude is three decimals. The forecast length, also called the lead time, is the time between the analysis time and the valid time. The AROME/Harmonie forecasts are performed with lead times of 3h, 6h, 12h, 15h, 18h, 21h and 24h and with analysis times of 00 UTC or 12 UTC. In this thesis, we use 3h lead time for 00 UTC to create forecasts for 03 UTC.

The observation data used in this thesis is for the Olkiluoto wind mast at the height of 60m. Olkiluoto is situated in southwestern Finland at a coastal area of the Gulf of Bothnia. The wind mast for which the forecasts are created is located on a top of a hill next to the sea, as displayed in Figure 3.4 (National Land Survey of Finland, 2014). The observation data includes the observed wind speed expressed in meters per second (m/s) and the observed wind direction in degrees (°). The accuracy of wind speed is one decimal while wind direction is given in integers. Observations are 10 minute averages of values measured just before the desired round moment, such as 00 UTC. The observed values for 03UTC are chosen in order to match with AROME/Harmonie data. Both the observation and the AROME/Harmonie data cover the period of 1st January – 17th August 2010. The days are expressed in Julian days to refer to data unambiguously.

Figure 3.4: The location of the Olkiluoto wind mast is pointed out with the arrow.

Chapter 4

Model and verification

4.1 Consistency of data and autocorrelation

Consistent data must have several characteristics. The observation and the initial AROME/Harmonie forecast data needs to have the same measurement time. In addition, systematical error related to AROME/Harmonie model height needs to be eliminated by choosing the best height available from AROME/Harmonie data. The AROME/Harmonie model height can differ from the height from where the observation data is taken, because AROME/Harmonie model does not illustrate perfectly forest, hills and other obstacles in the immediate surroundings of the wind mast that have an effect on wind speed. This characteristic of obstacles is called roughness (e.g. The Ministry of Employment and the Economy (Finland), 2013b). Roughness alters the wind profile since it affects both the wind direction and wind speed. The best matching height of AROME/Harmonie data is chosen by the method for least squares (e.g. Wilks, 2006). That is, the sum of the squared differences between corresponding AROME/Harmonie values and observation values is calculated for each model height of AROME/Harmonie data with the equation

$$S = \sum_{i=1}^{n} (f_i - o_i)^2,$$

where f_i are the AROME/Harmonie forecasts, o_i are the observed wind speed values and n is the amount of measurement times in the sample. The model height with the smallest sum of least squares is chosen.

Furthermore, autocorrelation of the observation data needs to be analysed. If the observation data is autocorrelated, the correlation should also be taken into the model.

Autocorrelation (e.g. Wilks, 2006) stands for the strength of the correlation among the consecutive values with a time lag between the values. Autocorrelation with no time lag represents the correlation of the value with itself, being always 1. Autocorrelation with one time lag, called the first lag, represents the correlation between the value and the next value. Autocorrelation with two time lags is called the second lag, and so on. Autocorrelation function consists of autocorrelations calculated for each value for several lags considering all the first lags together, all the second lags together etc. Autocorrelation function shows with which lags the values are correlated in the sample. If the correlation exceeds the confidence interval, for example 95% level, the values are correlated.

The model is created by using the chosen training sample and tested by using the chosen verification sample, also called the prediction sample. The training and verification data must be independent of each other (Hastie et al., 2009). Furthermore, the training data needs to be large enough so that single forecasts do not affect the results. If the observation values are autocorrelated, the training and verification data should include consecutive values. With data available from several years, the training data can involve data from several years and the verification data can involve data from one or more years. With a small amount of data, the data should be divided into K parts that are approximately of the same size, and the verification period should be in the middle of the data to represent the average conditions of all the data. (Hastie et al., 2009).If the observation values are not autocorrelated, the verification data can include data from all over the sample, for example every other day.

To test if the verification sample represents the average conditions, basic statistical scores — mean, standard deviation, maximum value and minimum value — can be compared for the observed wind speed values of both all the sample, the training sample and the verification sample. Mean is calculated with the equation

$$\mu = \frac{1}{n}\sum_{i=1}^{n} o_i,$$

where o_i are the observed wind speed values and n is the amount of measurement times in the sample. A standard deviation is calculated with the equation

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(o_i - \mu)^2}, \tag{4.1}$$

and the maximum value is simple the highest and the minimum value the lowest observed value in the sample.

4.2 Post-processing methods

Wilks (2006) crystallises the need for post-processing of NWP forecasts into three causes. Firstly, the NWP model homogenises the surface conditions of areas between grid points without taking small scale topography and other elements into account. Secondly, there may be systematic errors in NWP forecasts. Thirdly, the results of the NWP model are deterministic and do not cover the uncertainty of the forecasts. The AROME/Harmonie wind speed forecasts therefore need to be post-processed to better match the observations of the selected wind mast. The AROME/Harmonie forecast is only an average forecast for a 2.5 km $\times$ 2.5 km area and the actual wind speed is likely to deviate from the AROME/Harmonie value in different spots of the area. Especially in districts with low roughness — that is, no forests, hills or islands close to the mast — the real wind speed can be much higher than the average for the area.

Statistical post-processing methods based on Model Output Statistics (MOS) approach are commonly used in meteorology (Wilks, 2009). The MOS models develop deterministic NWP forecasts into more accurate forecasts, such as probability forecasts, by using statistical regression methods. In practice, NWP forecasts are used as predictor variables x_i in MOS regression equations. Wilks (2006) introduces two regression models that produce probabilistic predictand values: the regression estimation of event probabilities (REEP) and the logistic regression (LR). The difference between these models is the form of the link function. The REEP method uses probability directly as a predictand *f(x)* for multiple linear regression equations whereas the LR method uses a logistic link function for probability as a predictand to convert multiple non-linear regression equations into linear equations.

A probabilistic forecast stands for a probability that the deterministic value is above or below a certain threshold value. The disadvantage with the REEP model is that it gives also probabilities below 0 or above 1, which need to be considered as 0 or 1 respectively. The LR model gives probabilities between 0 and 1 and it is a widely used statistical post-processing method which Wilks (2006) and Wilks and Hamill (2007) proved to perform well compared to other post-processing methods. In the LR method, the threshold value is used as a predictor variable to form the forecast equations. The problem with the LR is that it does not give a full probability distribution because it calculates a separate equation for each threshold. Therefore, there is only a limited amount of thresholds for which the equations are formed and the equations may not be mutually consistent.

The extended logistic regression (ELR) is an extension of the logistic regression introduced by Wilks (2009). The ELR method involves a function of the thresh-

olds and forms only one forecast equation that takes all possible thresholds into consideration. In this thesis, we investigate if the ELR method improves the initial AROME/Harmonie forecast. The REEP method as well as the linear regression method are investigated as a reference models for the ELR to be able to evaluate the results. Wilks (2009) has showed that ELR improves forecast for precipitation and Messner et al. (2013) have investigated that the ELR can be used for forecasting wind speeds too.

4.2.1 Extended logistic regression

The extended logistic regression model produces probabilistic forecasts instead of deterministic values, such as wind speeds. In this thesis, we define that the probabilities p represent that the wind speed exceeds different threshold wind speeds. The choice does not affect the results since the probability that the wind speed is below the threshold is *1-p* and the verification results and error scores are the same as for probabilities p. With multiple thresholds the ELR gives a probability distribution of wind speed for each measurement time.

The LR method calculates a separate forecast equation for each threshold q. The probability p is a non-linear function of a regression equation *f(x)*. The LR model can be presented in the following form

$$p = \frac{e^{f(x)}}{1 + e^{f(x)}}, \tag{4.2}$$

where regression equation *f(x)* is a linear function of predictor variables x_i

$$f(x) = b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n.$$

By resolving the Equation (4.2) we get the logit form for the probability, also knows as log-odds

$$ln\left(\frac{p}{1-p}\right) = f(x). \tag{4.3}$$

In this form, the regression equation *f(x)* is linear with the logistic link function. A logistic function forms an S-shaped curve with probability p at the y-axis being restricted to an area of $0 < p < 1$.

In order to solve the intercept b_0 and coefficients b_i for the model, the regression equation *f(x)* is matched to the binary observation values. The deterministic observation values above the threshold are changed to 1 and observation values below the

threshold are changed to 0. The logistic regression function cannot be solved directly so an iterative maximum likelihood technique is used to solve the coefficients. Having solved the coefficients for the predictor variables x_i with the training data, it is possible to calculate the probability forecasts for the verification data from the Equation (4.3). The results achieved with different combinations of predictor variables are compared and the model with best verification results is chosen. Different verification methods are used to evaluate the probabilistic forecasts. Deterministic observation values of verification data are also changed to binary values to be able to calculate verification scores.

The LR creates a regression function for each threshold separately, which is troublesome. Additionally, the equations may not be mutually consistent. The ELR method has an advantage of producing one equation to cover all the thresholds q. In the ELR model, the function of thresholds *g(q)* is included in the Equation (4.2) and the ELR model gets the form

$$p(q) = ln\left(\frac{e^{f(x)+g(q)}}{1+e^{f(x)+g(q)}}\right).$$

The logit form is a linear function of both the regression function *f(x)* and the function of thresholds *g(q)*

$$ln\left(\frac{p(q)}{1-p(q)}\right) = f(x) + g(q). \tag{4.4}$$

Using the Equation (4.4), we can solve the coefficients for the predictor variables that work for all possible thresholds. The function of thresholds *g(q)* must be chosen together with model parameters. Although the calculations are carried out with a computer, it is not necessary to calculate values for infinitely many thresholds. The model is created with a set of threshold values and verified with a different set of thresholds. Regardless of the thresholds selected, the model coefficients should be the same.

Alternative predictor variables are formed from the data available. There are three different AROME/Harmonie forecasts available for different heights: wind speed, wind direction and temperature. These initial forecasts and functions of these forecasts can be used as predictor variables. The predictor variables chosen must have a physical justification. The AROME/Harmonie forecasts are numerical forecasts produced from the atmospheric equations and they already take into account wind speed, wind direction, temperature and other physical parameters. However, they are only average forecasts for the 2.5km $\times$ 2.5km area around the wind mast and post-processing can

improve the forecasts.

Since the ELR model will forecast wind speeds, the AROME/Harmonie forecast for wind speed is a rational alternative for the primary predictor x_1. Also the standard deviation of the AROME/Harmonie wind speed forecast is investigated as a predictor variable because it is used in other studies (e.g. Messner et al., 2013). Standard deviation variable is formed for each measurement time i by calculating the squared differences between the AROME/Harmonie wind speed forecast and the observed wind speed until the measurement time i in question

$$\sigma_i = \sqrt{\frac{1}{i}\sum_{k=1}^{i}(f_k - o_k)^2}, \tag{4.5}$$

where f_k are the AROME/Harmonie forecasts, o_k are the observed wind speed values and i is the measurement time in question. Denominator i is used instead of *i-1* so that the equation is valid for the value *i=1* as well.

Wind speed at the wind mast likely differs according to wind direction because roughness is different for wind blowing from different directions. Roughness depends on the obstacles right beside the wind mast. It may be smaller for wind blowing from a coastal direction assuming that there are less obstacles, such as islands. Due to roughness, wind speed can be higher or lower than the average AROME/Harmonie wind speed forecast from certain directions. It is reasonable to treat wind direction as a categorical variable. Otherwise, western directions would be given more weight than eastern directions without justification. There would also be a gap at 360°, although, the wind from the northwestern and northeastern directions could be equally strong. In this thesis, we choose to divide wind direction into offshore and onshore wind. It could possible to have more than two categories, but the sample must be large enough for each category to get statistically significant results.

Temperature T and temperature difference ΔT between temperatures at two heights are also experimented since they describe the physical characteristics of wind and there can be errors in the atmospheric equations of the AROME/Harmonie model.

Moreover, it is possible to form a variable called stability of the data available. Stability has an effect on the wind speed profile and it can be expressed with a gradient Richardson number Ri (Kaimal and Finnigan, 1994)

$$Ri = \frac{g}{\overline{\theta}}\frac{\partial\overline{\theta}/\partial z}{(\partial\overline{u}/\partial z)^2},$$

where g is the gravitational acceleration, θ is potential temperature, u is horizontal

wind velocity and z is height above the ground level. Potential temperature represents the temperature of the air parcel after it has been moved adiabatically (Holton, 1992)

$$\theta = T(p_s/p)^{R/c_p},$$

where T is the initial temperature of the air parcel, $p_s = 100$kPA is the standard pressure, p is the pressure of the air parcel, $R = 287$J/(K kg) is the gas constant for dry air and $c_p = 1004$J/(K kg) is the specific heat of dry air at constant pressure. Potential temperature can be approximated as follows

$$\begin{aligned} \theta_i &= T_i + \frac{g}{c_p} z \\ &\approx T_i + 0.01z, \end{aligned} \tag{4.6}$$

which is based on the dry adiabatic lapse rate (Holton, 1992)

$$\Gamma = -\frac{dT}{dz} = \frac{g}{c_p} \approx 0.01K/m.$$

Thus, the potential temperature increases the measured temperature T with approximately 0.01K/m. Potential temperature enables resolving of the Richardson number. The stratification of the atmosphere is stable for $Ri > 0$, unstable for $Ri < 0$ and neutral for $Ri = 0$. The flow changes from turbulent to laminar above a critical value. This critical value is 0.25 for inviscid flow (Kaimal and Finnigan, 1994), but there is evidence for turbulence existing at higher Ri values (e.g. Galperin et al., 2007). In this thesis we use Bulk Richardson number Ri_B which can be calculated with AROME/Harmonie temperature and wind speed values

$$Ri_B = \frac{g}{\overline{\theta}} \frac{\Delta\theta/\Delta z}{(\Delta u/\Delta z)^2} = \frac{g}{\overline{\theta}} \frac{\Delta\theta \Delta z}{(\Delta u)^2}. \tag{4.7}$$

The reference level is typically the ground level $z = 0$. Because stability as a predictor variable is experimental, the critical values of Ri do not hold good. The categories for the predictor variable Ri_B need to be chosen by trying alternative segmentations and selecting reasonable categories with enough data.

4.2.2 Reference models

The ELR and the reference models are compared to cross-validate the results. In this thesis, the regression estimation of event probabilities (REEP) model is the main reference model and the linear model (LM) the second reference model. The REEP method is an often used MOS model for creating probability forecasts (Wilks, 2009) and therefore comparable with the ELR. The results of the REEP model are verified the same way as the results of the ELR model because both models generate probability forecasts. The LM is a simple linear regression model and generates deterministic forecasts so probabilistic verification methods cannot be calculated. Consequently, we compare the basic error scores for the wind speed forecasts created with the ELR model, the REEP model and the LM in the end.

The REEP method is simply multiple linear regression where the predictand y_i is a probability and the predictor x_i is a continuous variable. In this thesis, the predictor x_i is the AROME/Harmonie forecast for wind speed. The REEP model is created with several thresholds and separate regression equations need to be formed for the thresholds. The formula for REEP model is

$$y_i = a_i + b_i x_i, \tag{4.8}$$

where a_i is the intercept and b_i is the coefficient for the predictor x_i. The observations are converted into binary values according to each threshold and used as the predictands y_i. The parameters a_i and b_i differ for each threshold q so the equation is solved several times for the same training data. The REEP Equation (4.8) for one threshold is exemplified in Figure 4.1.

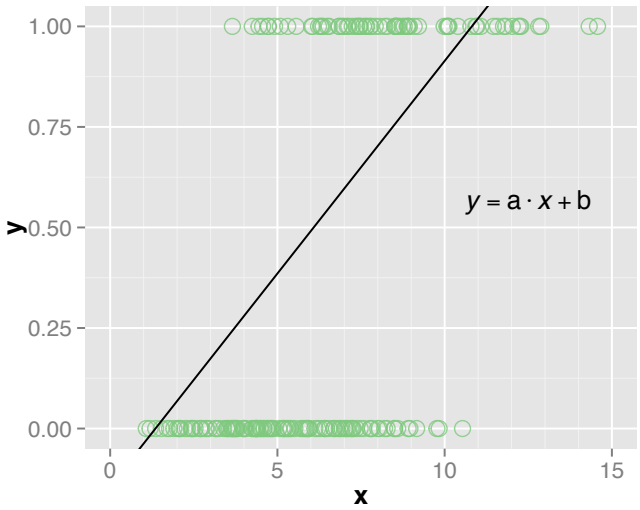


Figure 4.1: An example of a REEP regression function is illustrated with a black line.

Probability forecasts are calculated for each threshold q from Equation (4.8) by using the verification data and the solved coefficients a_i and b_i. The achieved predictand values y_i are directly the probabilities that the value exceeds the threshold q. The REEP equation is valid only between $0 \leq y_i \leq 1$. Predictand values $y_i < 0$ are changed to $y_i = 0$ and values $y_i > 1$ are changed to $y_i = 1$. Eventually, we get the forecasts for each threshold and can verify the results with binary observation values of the verification data.

The LM forecasts are calculated the same way as the REEP forecasts using Equation (4.8). The difference is that the predictand values are not probabilities but continuous deterministic values. Consequently, no threshold values are used and there is only one equation. When predicting wind speeds, the forecast values are directly in the form of m/s.

4.3 Statistical testing

Statistical testing methods are used to analyse the goodness of the model in the parametrical level. Significance of the parameters and goodness of the fit are basic statistical testing results. Significant parameters are included in the model whereas insignificant parameters are omitted.

Significance of parameters can be determined with the T-test. The T-test is a statistical test in which the value of a parameter is compared to a constant value (e.g. Laininen, 2000). When defining the significance of the predictor variables of the ELR model, the coefficients b_i of the predictor variables are compared to zero. First, normally distributed estimators of the coefficients $\hat{b_i}$ are formed with the predictor variables of the training data. The standard error of the estimators $SE[\hat{b_i}]$ is calculated as follows

$$SE[\hat{b_i}] = \frac{\sigma}{n},$$

where σ is the standard deviation for the predictor variables and n is the amount of measurement times in the training sample. The variance $SE^2[\hat{b_i}]$ responds the χ^2 distribution with the degrees of freedom $df = \upsilon$. The degrees of freedom df explain the number of values that are free to change.

The estimator values $\hat{b_i}$ are then compared to a constant value, in this thesis to the value 0. The test quantity t_0 is calculated as the standardised difference between the estimator $\hat{b_i}$ and the reference value

$$t_0 = \frac{\hat{b_i} - 0}{SE[\hat{b_i}]}.$$

P-values are calculated as a probability $P(T > t_0)$, where T is a t distributed random variable. The null hypothesis H_0 is that the coefficients b_i of the predictor variables i are zero, H_0: $b_i = 0$. The null hypothesis is abandoned if the P-value ≤ 0.05, that is, with 5% significance level (e.g. Laininen, 2000). If the null hypothesis is abandoned, the alternative hypothesis H_1 takes effect, H_1: $b_i \neq 0$, which means that the model parameter is significant and should be included in the model. Therefore, the P-value is the probability that the null hypothesis is true, in other words, the probability that the same results are obtained by change. The smaller the P-value, the more significant the parameter.

In addition to the significance of the model parameters, the goodness of the fit of the model is calculated to assess how well the model fits the observations. The goodness of the fit of the model, R^2 (e.g. Laininen, 2000), is calculated from the following equation

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

where SST is the total sum of squares describing the overall variability of the predictand, SSR is the regression sum of squares, in other words, the sum of squared

differences between the regression predictions and the sample mean and SSE is the sum of squared errors, also called the residual sum of squares. The total sum of squares SST is calculated from equation

$$SST = \sum_{i=1}^{n} (o_i - \bar{o})^2,$$

where o_i are the observed values and $\bar{o}$ is the mean of the observations. The regression sum of squares SSR is calculated from equation

$$SSR = \sum_{i=1}^{n} (y_i - \bar{o})^2,$$

where y_i are the predicted values. The residual sum of squares SSE is calculated with the observed values o_i and predicted values y_i from equation

$$SSE = \sum_{i=1}^{n} (o_i - y_i)^2.$$

The total sum of squares SST equals the regression sum of squares SSR and the residual sum of squares SSE

$$SST = SSR + SSE.$$

The goodness of the fit R^2 is therefore the proportion of variance of the predictand described by the regression model. The goodness of the fit value range between 0 and 1 so that the closer R^2 is to 1, the better the regression. The goodness of fit can be used to compare different models.

4.4 Verification

Verification is a procedure used in atmospheric science for evaluating the quality of the forecast with numerical measures such as those described by Jolliffe and Stephenson (2003). Verification scores give information of the advantages and disadvantages of the model; whether the forecasts are accurate and reliable, and the model able to discriminate events with different characteristics. The sample climatology, the average of the observed values, is often used as a reference to evaluate the performance of the new model. The verification scores for the forecasts can also be compared to assess the competing models from a wider perspective.

There are various verification methods to describe characteristics of forecasts. In this thesis, verification methods for probability forecasts are used to choose the best ELR model and to compare the chosen ELR model with the REEP model. The results of the verification methods and statistical testing are evaluated jointly because there are no simple priority rules to tell which results are more important. Consequently, we select a model for which both the verification and statistical testing scores give good results considering a reasonable scale. We select the significant amount of significant numbers to be taken into account and so decide which change in the score has a noticeable effect. The sensitivity for the verification scores and the statistical results is tested by verifying subsets of the sample with the bootstrap method.

In order to compare the ELR and the REEP forecasts with the LM forecasts and the initial AROME/Harmonie forecasts, the probabilistic forecasts are altered into wind speed values. The basic verification scores are used to examine the error for the wind speed forecasts and the observations. Moreover, the graphical analysis of the error quantiles and the fit of the wind speed forecasts and the observations illustrate the size and the significance of the errors.

The following verification methods are described both by Jolliffe and Stephenson (2003), The Centre for Australian Weather and Climate Research (2013) and verification learning modules of Eumetcal (Nurmi, 2014). Basic verification scores are used to analyse the errors of deterministic forecasts and other verification scores are used to analyse the probability forecasts.

4.4.1 Basic verification scores

Basic measures for verification are the mean error (ME), the mean absolute error (MAE) and the root mean square error (RMSE). The mean error, also known as the bias, is the average error

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i),$$

where n is the amount of forecasts, y_i is the deterministic forecast and o_i is the deterministic observation. In this context, a deterministic value refers to a wind speed. Because errors can be either positive or negative they probably compensate each other. That is taken into consideration with the mean absolute error which regards all errors as positive. The mean absolute error is the average of the absolute values of the errors

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - o_i|.$$

The mean absolute error describes the relative magnitude of the error, that is, the average distance between the forecasts and the observations. The third basic verification score for errors is the root mean squared error. It is calculated as the root of mean squared error (MSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2}$$

The root mean square error gives more weight to large errors, resulting from the squaring of the errors before taking the average. Therefore the *RMSE* is advantageous if large errors are unwanted. In this thesis, *RMSE* matters more than *ME* or *MAE* because large errors have a significant effect on the yield of the wind power plant. It is also practical to calculate the absolute maximum and minimum error. Nonetheless, the closer *ME*, *MAE* and *RMSE* are to zero, the smaller the error.

4.4.2 Brier score

The Brier score (BS) is a verification method to calculate the accuracy of the forecast. It is the mean squared error of the probabilistic forecasts computed over the verification sample. A probability forecast y_i gets a value between 0 and 1 whereas an observation o_i gets a binary value 1 if the event occurs and value 0 if it does not. The squared error is therefore at maximum 1. The Brier score is the mean of the squared errors so it ranges between 0 and 1, perfect score being 0. The Brier score is calculated with the following formula

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - o_i)^2,$$

where y_i is the probabilistic forecast, o_i is the probabilistic observation and n is the amount of forecasts. Brier score can be divided into three components

$$BS = \frac{1}{n}\sum_{k=1}^{K} N_k(y_k - \bar{o_k})^2 - \frac{1}{n}\sum_{k=1}^{K} N_k(\bar{o_k} - \bar{o})^2 + \bar{o}(1-\bar{o})$$
$$= \text{"}Reliability\text{"} - \text{"}Resolution\text{"} + \text{"}Uncertainty\text{"}.$$

The first component represents the reliability of the probabilistic forecasts, the second component represents the resolution and the third component represents the uncertainty. In this decomposition of Brier score, the probabilistic forecasts y are divided into K bins, which means that forecasts are put into groups according to their value. For example, in decile bins the groups are 0 – 10%, 10 – 20%, 20 – 30% and so on. The symbol $\bar{o_k}$ stands for the observed relative frequency of the event in a bin k and the symbol $\bar{o}$ stands for the overall sample relative frequency also known as the sample climatological frequency. N_k is the amount of forecasts in a bin k.

The reliability term is the mean of weighted and squared differences between the binned probabilities and the observed relative frequencies. Therefore, it is the conditional bias of the forecasts. The smaller the reliability, the better the score. With zero reliability, the sample is perfectly reliable.

The resolution term is the mean of weighted and squared differences between the observed relative frequencies and the sample climatological frequency. It measures the capability of the model to separate the events into different types. Because there is a negative sign before the resolution term, the best resolution score is as big as possible.

The uncertainty is the Brier score for the sample climatology. Uncertainty depends only on the observation values so the Brier scores for different samples should not be compared. Uncertainty gets a maximum value of 0.25 with sample climatology of 0.5 and a minimum value of 0 with sample climatology of 0 or 1.

4.4.3 Brier skill score

The Brier skill score (BSS) measures the skill of the model by comparing the accuracy of the model to the accuracy of the reference model. Thus, it proves if the model is better or worse than the reference model. A reference model is either the sample or long-term climatology. The Brier skill score is calculated as the difference between the Brier scores for the model and for the reference model divided by the difference between the Brier scores for a perfect model and for the reference model. The denominator represents the maximal improvement that the model can attain. The formula for the Brier skill score is

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}}.$$

The Brier skill score ranges between -∞ and 1. The best Brier skill score is 1 implicating that $BS = 0$. A positive BSS stands for a more accurate model than the reference and a negative BSS stands for a worse model than the reference. If the BSS value is 0, the model has no skill compared to the reference model.

The Brier skill score does not give reliable results if the sample is too small or if the sample BSS is too low. A low sample climatology, which is common for rare events, indicates that the denominator of BSS is small and therefore the skill score is unstable. The use of the long-term climatology is better because it leads to a more representative outcome and a better BSS. Unfortunately, the long-term climatology is not always available.

4.4.4 Reliability

Reliability shows how well the model works in different circumstances. It represents the ability of the model to create different kind of forecasts for different types of events, for example, for low and for high wind speeds. The error may be nonlinear, which should be taken into account when predicting values.

In a reliability diagram forecast probabilities are divided into bins according to their values. The bins are plotted against the relative observed frequency of the event for the corresponding cases. The relative observed frequency stands for the relative number of occurrences of the event for the probability bin. Forecast bins can be for example deciles. If there is more than one threshold, all the forecasts of all the thresholds are treated together for the bins and relative observed frequencies.

A reliability diagram is illustrated with points connected with lines in Figure 4.2 (The Centre for Australian Weather and Climate Research, 2013). For a perfectly reliable forecast, the relative observed frequencies lie in the diagonal. Points above the diagonal mean that the model is under-forecasting so the event occurs more often than the model forecasts. Respectively points below the diagonal mean that the model is over-forecasting. The horizontal line in the diagram is the climatological frequency that denotes the relative observed frequency for the whole sample. The line halfway between the diagonal and the horizontal line indicate points that have no skill.

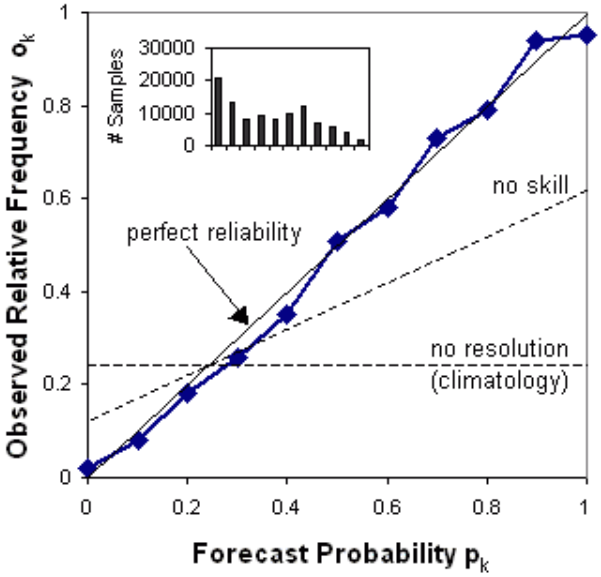


Figure 4.2: An example of a reliability diagram. The probability of detection stands for the hit rate.

A sharpness diagram is a small histogram, which illustrates how many forecasts are included in each bin. It is usually included in the reliability diagram. In Figure 4.2 it is in the top left corner. Sharpness represents the variance of the forecast probabilities. Perfect sharpness can be reached only with a categorical yes/no -forecast for which one half of the forecasts have value 0 and the other half have value 1. Sharpness is good for a probabilistic forecast of a continuous variable if there is a U-shape in the sharpness diagram. It refers to a large amount of cases with a high or a low probability.

Reliability is the mean of the weighted and squared differences between the binned probabilities and the observed relative frequencies. In the reliability diagram, it is the mean squared vertical distances between the observed relative frequency points and the diagonal. Resolution is the mean of the weighted and squared differences between the reliability points and the sample climatological frequency. In the reliability diagram, it is the mean squared vertical distances between the reliability points and the horizontal line.

4.4.5 Relative operating characteristics

Relative operating characteristics (ROC) is a measure of discrimination. Binary observations, indicating that the event either occurred or not, are divided into two categories: occurrences and non-occurrences. The distribution of the probability forecasts is formed for both the categories separately. High probabilities should interconnect with occurrences and low probabilities with non-occurrences to minimise the overlap of these two conditional distributions. The less overlap, the better the model discriminates events.

The distributions are examined with probability thresholds for forecasts. A probability threshold means that probabilistic forecasts with a value above the threshold are taken into account to calculate the hit and false alarm rate for that threshold. In a ROC diagram, several probability thresholds cover the range of probabilities from 0 to 1. Depending on the threshold, the forecast get a categorical value 'yes' or 'no' to display if the event occurred or not. The forecasts with a value above the threshold get a categorical value 'yes' and forecasts with a value below the threshold get a categorical value 'no'. Hits are the occurrences of the event for forecasts with a categorical value 'yes'. Hit rate is the number of hits divided by the total amount of occurrences. False alarms are all the rest of the cases, all the non-occurrences, for forecasts with a categorical value 'yes' since they forecasted incorrectly that the event would occur. False alarm rate is the amount of false alarms divided by the total amount of non-occurrences. Hit and false alarm rates are consequently relative amounts. A hit rate – false alarm rate value pair for each probability threshold used is plotted in a two-dimensional ROC diagram in Figure 4.3 (The Centre for Australian Weather and Climate Research, 2013).

The ROC diagram in Figure 4.3 shows that a small probability threshold have both a high hit rate and a high false alarm rate because it considers all the forecast above that threshold as 'yes' events. For large probability thresholds both the rates are small. The points for no warnings (zero hits and zero alarms) and for always warnings (100% hits and 100% alarms) are included in the diagram. All the points in the diagram are connected. Provided that the sample is big enough, the fitted curve can be smooth, otherwise, the points can be connected with straight lines.

The area under the ROC curve measures the discrimination level of the model. The bigger the area, the better the model. The top left corner stands for 100% hits and so the curve should be close to that point to maximise the area under the curve. The diagonal, for which the hit rate is exactly the same as the false alarm rate, stands for no discrimination and therefore no skill.

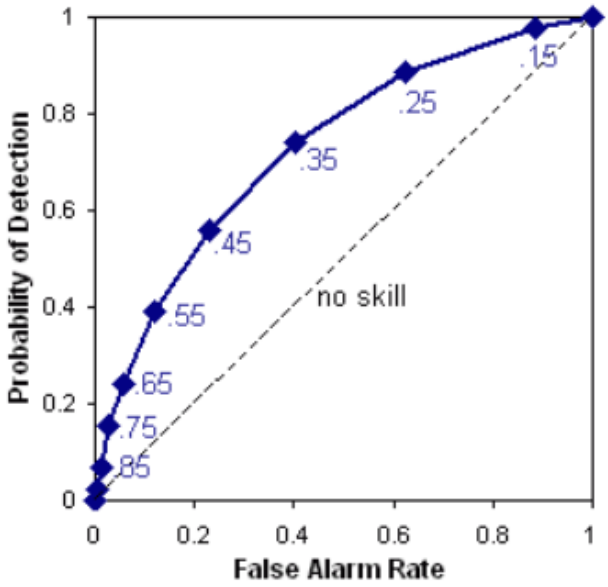


Figure 4.3: An example of a ROC diagram.

The ROC can be considered as a measure of usefulness because it describes the performance of forecasts. The thresholds can be used in decision making since we know the critical hit and false alarm rates for events. The weakness of the ROC is that it is insensitive to calibration of probability forecasts, such as adding a constant to probabilities, because it does not affect the shape of distributions. Therefore, the ROC should not be used alone but together with reliability.

4.5 Bootstrap method

The bootstrap is a method to estimate the statistical sensitivity (e.g. Hastie et al., 2009). The method is used to demonstrate similar possible situations when there is not enough data available (Wilks, 2006). The idea of bootstrapping is to take a random subset of the data several times, e.g. 100 times, and create a model with each of the datasets. Consequently, we get several possible models with some variation. According to Wilks (2006), bootstrapping is generally made with resampling with replacement, that is, the same data values can be chosen several times because they are not excluded from the sample before choosing next values. Respectively, some data values are not chosen at all. Resampling without replacement is called permutation.

Chapter 5

Results and discussion

5.1 Data choices

The AROME/Harmonie model height is chosen by assessing the least squares. The least squares for the AROME/Harmonie data for heights 30m – 60m and the observation data are examined. The purpose is to choose the AROME/Harmonie height for which the forecasts match best with the observations and to eliminate the possible systematic error in the model height caused by roughness. The least squares round to integers are shown in Table 5.1.

AROME/Harmonie model height [m]	**Sum of least squares [m^2/s^2]**
60	1393
50	1254
47	1216
30	1045

Table 5.1: The least squares for the AROME/Harmonie data for heights 30m –60 m and the observation data at 60m.

The sum of the least squares is smallest for the AROME/Harmonie forecasts at 30m. Hence, the AROME/Harmonie values at 30m match best to the observed values at 60m. The fit of the observations and the AROME/Harmonie forecasts for heights 30m and 60m is illustrated in Figure 5.1. The closer the values are to the diagonal, the better the fit. Figure 5.1 shows that the systematic error in the AROME/Harmonie model forecasts is compensated by choosing the best fitting model height but there is still some error in the forecasts.

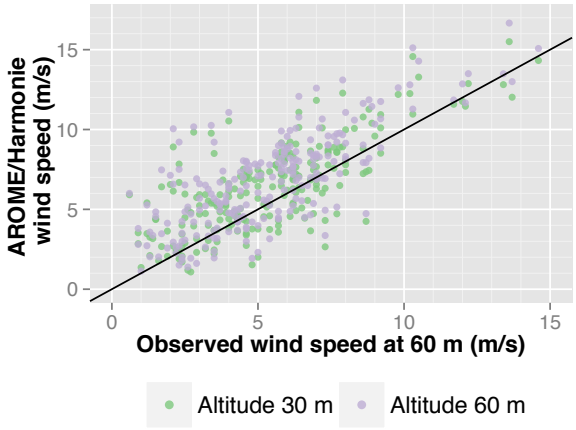


Figure 5.1: The observed wind speed values versus the AROME/Harmonie wind speed values at heights 30m and 60m.

Autocorrelation of the observed wind speed values has to be examined in order to know if the consecutive wind speed values are correlated. If there is correlation, it needs to be taken into account in the model. The autocorrelation function of the observed wind speeds created for all the data is presented in Figure 5.2. The time lag for the forecasts at 03 UTC is one day. The dashed blue line illustrates the 95% confidence interval.

Figure 5.2 shows that the first lag exceeds the 95% confidence interval. Thus, the observations are autocorrelated with the first proceeding values. Other time lags are inside the confidence interval so they are not autocorrelated. Consequently, the training and the verification data subsets have to include consecutive values. The whole data sample includes 229 Julian days. We choose to use cross-validation for scarce data and divide the sample into K roughly equally sized parts. We divide the data into 8 subsets so that each month represents one subset. Substantially much data is used for creating the model and less data for the verification. April is chosen to be the verification period because it represents well the average conditions of all the data both ocularly and according to statistics. April includes Julian days 91 – 120 with the first day being 1$^{\text{st}}$ April and the last day being 30$^{\text{th}}$ April. The training period includes Julian days 1 – 90 with the first day being 1$^{\text{st}}$ January and the last day being 31$^{\text{st}}$ March and Julian days 121 – 229 with the first day being 1$^{\text{st}}$ May and the last day being 17$^{\text{th}}$ August.

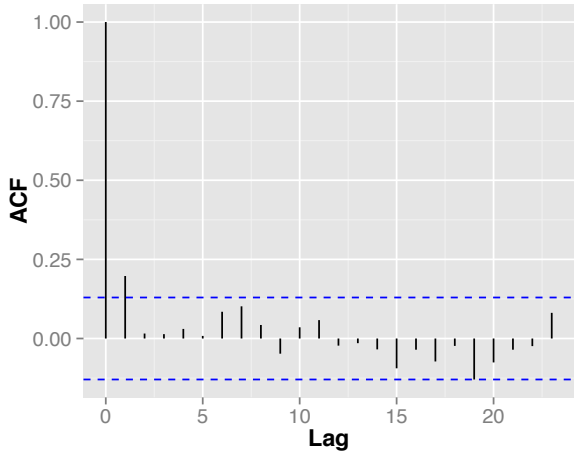
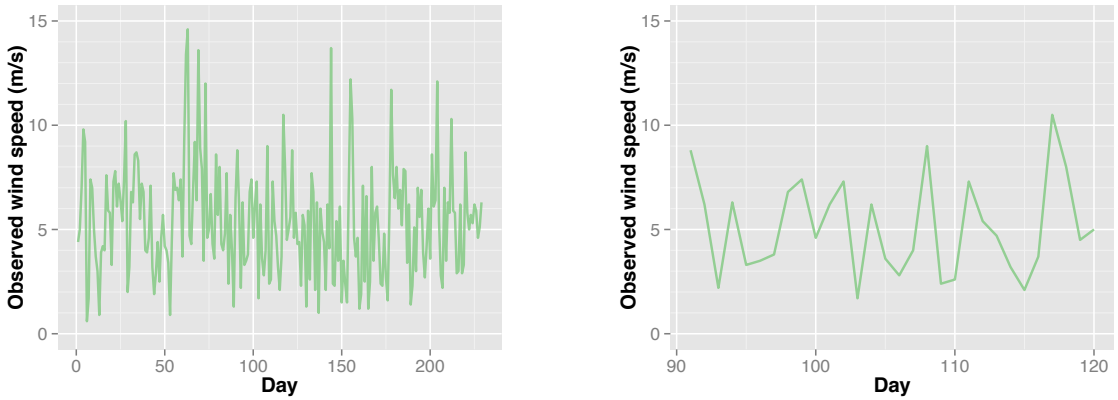


Figure 5.2: The autocorrelation of the observed wind speed values.

The basic statistical scores — mean μ, standard deviation σ, maximum value and minimum value — are calculated for the chosen verification data to show that it represents the average conditions. Mean and standard deviation are calculated from the Equations 4.1 and 4.1 respectively. The basic scores are presented in Table 5.2 and the wind speeds are plotted over time in Figure 5.3.

Data	μ **[m/s]**	σ **[m/s]**	**Max [m/s]**	**Min [m/s]**
All the data	5.4	2.6	14.6	0.6
April	5.1	2.3	10.5	1.7

Table 5.2: The basic statistical scores for the training and the verification data.



(a) The observed wind speed values for all the data.

(b) The observed wind speed values for April.

Figure 5.3: The observed wind speed curves.

The basic statistical scores and graphical evaluation suggest that the verification sample represents the average conditions well. High wind speed values are missing in the verification sample but there are only few high wind speeds in the whole sample likewise. Therefore, it would not be possible to model high wind speeds extensively with any other verification period either.

5.2 Extended logistic regression model

The ELR model is created with the programming language R. First, the predictor variables chosen to be experimented are extracted from the AROME/Harmonie data for all Julian days including both the training sample and the verification sample. After the selection of the alternative predictor variables, the model coefficients are calculated the with the *BAM* function of R which fits the observations and the predictor variables by using a logit link function as in Equation 4.4. The predictor variables chosen are wind speed, standard deviation of wind speed, wind direction, temperature, temperature difference and stability.

5.2.1 Predictor variables

The AROME/Harmonie wind speed forecast is used as the primary predictor x_1. The standard deviation is calculated for the AROME/Harmonie wind speed data according to Equation 4.5 and experimented as the predictor variable x_2.

Wind direction is chosen to be a categorial variable x_3. The wind direction predictor variable is created by classifying the data into two categories, offshore and onshore wind, according to the AROME/Harmonie wind direction forecast. Figure 3.4 shows that the segmentation is not obvious. Consequently, different boundary values for the categories are experimented. The categories are given different values to separate them from each other and the coefficient for these values is defined by the BAM function. We choose to give the offshore category value 1 and the onshore category value 0. The choice of these values does not affect the results as long as there are only two categories since the coefficient is included in the model as well. This is tested empirically with different values, such as a value 100 for the offshore and 0 for the onshore category, as well as 100 for the offshore and -100 for the onshore category.

The stability variable x_4 is created as a Bulk Richardson number Ri_B presented by Equation 4.7. The AROME/Harmonie data does not exist at the ground level $z = 0$. Thus, the potential temperatures θ_{i} are calculated from Equation 4.6 for the AROME/Harmonie temperature forecasts at heights 30m and 47m. The average potential temperature $\overline{\theta}$ is calculated as the average of these two potential temperatures. The AROME/Harmonie wind speed forecasts u are also taken from heights at 30m and 47m. The stability variable is formed for each Julian day. The stability variable can get huge positive values that have a great impact on the model coefficients in case the model weights the stability values directly. Therefore, also the stability values are categorised.

In this thesis, only two stability categories are used. Several alternative stability categories would make the model complicated since there are numerous alternatives to give a value for each category and the choice of values could not be easily justified. The model is simpler with only two stability categories, because the choice of the values given to two categories does not affect the results. Thus, the logical category choice is to divide stability values only into unstable ($Ri_B < 0$) and stable ($Ri_B > 0$) atmosphere, instead of several stable categories according to the magnitude of stability or an interval close to zero describing neutral atmosphere. There is a wide physical difference between stable an unstable situations so the choice of categories is reasonable. There are 215 values in the stable and 14 values in the unstable category. The number of values in the category for unstable atmosphere ($Ri_B < 0$) is small, which may not bring out the influence of stability well. However, the forecasts are done for night-time (03 UTC) so it is reasonable that there are much more stable situations than unstable. Therefore, we assume that 14 values are enough to make this category statistically significant.

Moreover, the AROME/Harmonie temperature forecasts T at 30m is used directly as a predictor variable x_5. Temperature difference between the AROME/Harmonie

temperature forecasts at 30m and 47m, $\Delta T = T_{47}$ - T_{30}, is used as another predictor variable x_6. The AROME/Harmonie forecasts at 47m are chosen as a reference point because forecasts at the ground level are not available.

5.2.2 Alternative models

The ELR model includes several wind speed thresholds q for which the probability values are calculated. The ELR model will cover all possible thresholds by using a function of the thresholds *g(q)* so the choice of threshold values q can differ for the training and the prediction data. In this thesis, threshold values q = 0, 4, 8, 12, 15, 20 are used for the training data and threshold values q = 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24 are used for the prediction data. The corresponding binary predictand values are calculated for the observation data o_i with each of the chosen threshold values q. The observed wind speed values above the threshold get a binary value 1 and the observed wind speed values below the threshold get a binary value 0. The ELR model as in Equation 4.4 experimented is

$$ln\left(\frac{p(q)}{1-p(q)}\right) = f(x) + g(q) = b_0 + g(q) + b_1x_1 + b_2x_2 + ... + b_nx_n,$$

where b_i are the coefficients, x_i are the predictor variables tested and n is the amount of predictor variables. The choice of the predictor variables and *g(q)* is empirical. The functions *g(q)* experimented are $g(q) = c_1\sqrt{q}$ and $g(q) = c_2q$, where c_i is the coefficient. Both of these functions are used in the literature Wilks (2009). As shown later in the thesis, the square root of threshold works better than thresholds alone. Therefore, more results are shown with the square root of thresholds.

Nine models involving different predictor variable combinations are tested: A, B, C, D, E, F, G, H and I. The predictor variable combinations used in each model are presented in Table 5.3. The primary predictor variable x_1 included in every model is the AROME/Harmonie wind speed forecast, the predictor variable x_2 is the standard deviation of the AROME/Harmonie wind speed forecasts, the predictor variable x_3 is the AROME/Harmonie wind direction forecast category, the predictor variable x_4 is the stability category created with the AROME/Harmonie forecasts, the predictor variable x_5 is the AROME/Harmonie temperature forecast T and the predictor variable x_6 is the AROME/Harmonie temperature difference ΔT. The primary predictor variable is naturally used in each model. Models A and I differ only according to the function of threshold *g(q)*. The model A involves the square root of the threshold whereas the model I involves the threshold directly.

	g(q)	x_1	x_2	x_3	x_4	x_5	x_6
Model	Function of threshold	Standard deviation	Wind speed	Wind direction category	Stability category	T	ΔT
A	$\sqrt{q}$	x	-	-	-	-	-
B	$\sqrt{q}$	x	x	-	-	-	-
C	$\sqrt{q}$	x	-	x	-	-	-
D	$\sqrt{q}$	x	-	-	x	-	-
E	$\sqrt{q}$	x	-	-	-	x	-
F	$\sqrt{q}$	x	-	-	-	-	x
G	$\sqrt{q}$	x	-	x	x	-	-
H	$\sqrt{q}$	x	-	x	-	x	-
I	q	x	-	-	-	-	-

Table 5.3: Predictor variables involved in each model are marked with x.

The *BAM* function of R enables the use of prior weights for the data, such as the AROME/Harmonie wind speed or the standard deviation. The data can also be weighted equally, which corresponds to a situation with no weighting. We choose to examine five different weights: W_0 for equal weighting, W_1 for the AROME/Harmonie wind speed, W_2 for the standard deviation of the AROME/Harmonie wind speed, W_3 for the square of the AROME/Harmonie wind speed and W_4 for the cube of the AROME/Harmonie wind speed. The model coefficients are calculated with the *BAM* function for all of the models presented in Table 5.3 and the five prior weights by using the training data.

5.3 Verification of extended logistic regression models

Verification scores for the probability forecasts and statistical testing results for the models are analysed to discover the ELR model with the best parameter combination. The *predict* function of R is used with the type *response* to solve the probability forecasts for the prediction period with the ELR models A – I. The probability forecasts are then verified against the corresponding binary observation values with the *verify* function of R. The *verify* function gives following verification scores: Brier score, Brier skill score, reliability, resolution and uncertainty. The sample climatology is used as a reference model. The *verify* package of R also enables plotting of the reliability and the ROC diagrams as well as calculating the ROC area. The *summary.gam* command

computes the P-values for the predictor variables and the goodness of the fit value by using Wald tests that are similar to the T-test.

The BSS is considered as the most significant verification score because it describes if the model works better than the sample climatology. The BS describing the accuracy of the forecasts is considered as the second important verification score. Reliability is included in the BS so the analysis of the reliability diagram is used only in the further evaluation of the best models together with the ROC diagram. Resolution and uncertainty depend only on the observation data that is the same for all the alternative models so they are not analysed. The models involving P-values for the parameters that are greater than 0.05 are omitted.

P-values and the goodness of the fit measure the goodness of the model in a parametrical level whereas the BSS and the BS measure the goodness of the forecasts in an independent sample. Therefore, it is reasonable to assess both the statistical testing scores and the verification scores together and discuss if the differences in the decimals are statistically significant. The accuracy of three significant numbers are used for the verification scores and the goodness of the fit to separate the models from each other. The accuracy of two significant numbers is used for the P-values not to exclude models at the boundary limit since the amount of data is relatively small and changes in the third decimal would not be credible. The sensitivity of the verification scores and the statistical testing results are compared for the best ELR models in the further evaluation.

5.3.1 Choice of wind direction categories

The best boundary values for the wind direction categories are determined by comparing the verification results of different category alternatives. The AROME/Harmonie wind speed x_1 and the wind direction category x_3 are only used as predictor variables to ignore the effects of other variables. Thus, the model C in Table 5.3 with different prior weights is used. The categorical choice should not be dependent on single values so the boundary values are at 10° intervals. Table A.1 in the Appendix A shows that there are at least 50 values in each category evaluated, which is enough values to make the tested category choices statistically significant.

The verification scores and statistical testing results for the model C with different wind direction categories are divided according to prior weights in Tables A.2, A.3, A.4, A.5 and A.6 in the Appendix A. P-values are included to determine with which category choices the wind direction variable is acceptable. Unacceptable category choices are indicated with red colour. The best boundary values for the categories differ ac-

cording to the prior weight. The small differences in the verification results are likely due to the small amount of data. Consequently, we choose the wind direction segmentation by assessing which category alternative works best for all prior weights on average. The BSS values range between 0.792 and 0.812, the BS values range between 0.338 and 0.376 and the goodness of the fit values range between 0.817 and 0.834.

The onshore category is chosen to be 0° – 240° and the offshore category 240° – 360°. These categories work best for the prior weight W_2 and they are among the best categories for all the other prior weights. The verification scores with the onshore category 0° – 250° are the best or the second best for the prior weights W_0, W_1 and W_4 but the scores are among the poorest for the prior weights W_2 and W_3. Thus, the onshore category 0° – 250° is worse than the onshore category 0° – 240° although wind direction category is an acceptable parameter for both of these models with all the prior weights. Therefore, we conclude that onshore category 0° – 240° is on average the best choice for all the prior weights.

5.3.2 Verification scores and statistical testing results

Verification scores and statistical testing results are calculated for all models presented in Table 5.3. The results are shown in Tables 5.4, 5.5, 5.6, 5.7 and 5.8. Rows marked with red colour stand for models involving at least one parameter that is not acceptable with a P-value greater than 0.05. Thus, models in red are not taken into consideration when choosing the best model. The best models for each prior weight are indicated with green.

Table 5.4 shows that only models A, B, C and I are acceptable with equal weighting. The best verification scores are obviously achieved with the model C. The goodness of the fit for the model C is only 0.003 smaller than that for the model B. Thus, including wind direction variable in the ELR model improves forecast quality with the prior weight W_0.

Model	BS	BSS	R^2
A	0.0368	0.796	0.818
B	0.0370	0.795	0.822
C	0.0353	0.804	0.819
D	0.0366	0.797	0.818
E	0.0371	0.794	0.821
F	0.0369	0.795	0.822
G	0.0354	0.803	0.819
H	0.0361	0.800	0.821
I	0.0366	0.797	0.815

Table 5.4: Verification scores and the goodness of the fit with the prior weight W_0.

Table 5.5 shows that models F and G are not acceptable with the AROME/Harmonie wind speed as the prior weight because the temperature difference variable (model F) and the stability variable (model G) have P-values greater than 0.05. Models C and H have clearly the best verification scores and the goodness of the fit. The verification scores are exactly the same for these models with three decimal accuracy. Therefore, the wind direction variable in the model C and the wind direction variable and the temperature variable in the model H improve the forecast quality notably compared to the model A with the AROME/Harmonie wind speed alone as the predictor variable. However, it is not reasonable to include temperature as an extra parameter in the ELR model because it does not give any additional value since the goodness of the fit is only 0.001 better than that for the model C. Consequently, the model C is the best choice with the prior weight W_1.

Model	BS	BSS	R^2
A	0.0354	0.804	0.815
B	0.0365	0.798	0.818
C	0.0341	0.811	0.822
D	0.0357	0.802	0.816
E	0.0367	0.797	0.817
F	0.0361	0.799	0.815
G	0.0343	0.810	0.822
H	0.0341	0.811	0.823
I	0.0359	0.801	0.809

Table 5.5: Verification scores and the goodness of the fit with the prior weight W_1.

Table 5.6 shows that models D, E, G and H are not acceptable with the standard

deviation of the AROME/Harmonie wind speed as the prior weight because the stability variable (models D and G) and the temperature variable (models E and H) have P-values greater than 0.05. The verification scores are obviously best for the model C with the wind direction as the predictor variable. The goodness of the fit for the model C is only 0.04 smaller than that for the model B and 0.03 smaller that that for the model F. The model C is therefore superior to other models also the prior weight W_2.

Model	**BS**	**BSS**	**R^2**
A	0.0366	0.797	0.817
B	0.0370	0.795	0.822
C	0.0344	0.809	0.818
D	0.0364	0.798	0.817
E	0.0374	0.792	0.819
F	0.0370	0.795	0.821
G	0.0342	0.810	0.818
H	0.0358	0.801	0.820
I	0.0368	0.796	0.814

Table 5.6: Verification scores and the goodness of the fit with the prior weight W_2.

Table 5.7 shows that all the models are acceptable with the square of the initial AROME/Harmonie wind speed as the prior weight because the P-values are zero with two decimal accuracy for all the predictor variables. The best verification scores and the goodness of the fit are achieved with the model G that involves both wind direction and stability as the predictor variables. The goodness of the fit for the model G is only 0.001 smaller than that for the model H and the verification scores are clearly superior for the model G. The verification scores and the goodness of the fit for the model C are good but slightly outperformed by those for the model G.

Model	BS	BSS	R^2
A	0.0349	0.806	0.809
B	0.0358	0.801	0.811
C	0.0348	0.807	0.826
D	0.0348	0.807	0.81
E	0.0354	0.804	0.812
F	0.0344	0.809	0.809
G	0.0344	0.809	0.827
H	0.0354	0.803	0.828
I	0.0356	0.802	0.803

Table 5.7: Verification scores and the goodness of the fit with the prior weight W_3.

Table 5.8 shows that all the models are acceptable with the cube of AROME/Harmonie wind speed as the prior weight because the P-values are zero with two decimal accuracy for all the predictor variables. None of the models have both the best verification scores and the best goodness of the fit. The model E has the very best verification scores so it is included in further analysis regardless the mediocre goodness of the fit. The models C, G and H have the best goodness of the fit. Since the models have approximately the same verification scores and the goodness of the fit, it is not reasonable to include stability or temperature as extra parameters in the ELR model because they do not give any additional value compared to situations without them. Consequently, only the model C with the wind direction as the predictor variable is included in the further analysis together with model E with temperature as the predictor variable.

Model	BS	BSS	R^2
A	0.0350	0.806	0.801
B	0.0352	0.805	0.803
C	0.0355	0.803	0.830
D	0.0357	0.802	0.803
E	0.0343	0.810	0.806
F	0.0350	0.806	0.803
G	0.0357	0.802	0.831
H	0.0355	0.803	0.832
I	0.0382	0.788	0.797

Table 5.8: Verification scores and the goodness of the fit with the prior weight W_4.

To infer, the model C with the wind direction variable is the best model with all the prior weights except W_3. In addition, the model C is acceptable with all the prior

weights. The model G with the wind direction and the stability variables together is the best model with the prior weight W_3 but it is acceptable only with the prior weights W_3 and W_4. Likewise, the model E with the temperature variable is one of the best models with the prior weight W_4 but it is acceptable only with the prior weights W_1, W_3 and W_4. The model A is superior to or equally good compared to the model I with all the prior weights. The use of the square root of the threshold as a model parameter in other models is therefore justified. The best results for each prior weight are gathered into Table 5.9.

Prior weight	**Model**	**BS**	**BSS**	$\mathbf{R^2}$
W_0	C	0.0353	0.804	0.819
W_1	C	0.0341	0.811	0.822
W_2	C	0.0344	0.809	0.818
W_3	G	0.0344	0.809	0.827
W_4	C	0.0355	0.803	0.830
W_4	E	0.0343	0.810	0.806

Table 5.9: The best verification scores and the goodness of the fit with each prior weight.

Table 5.9 shows that there is no single model to stand out from the rest. The model C with the prior weight W_1 and the model G with the prior weight W_3 are the best when taking into account both the verification scores and the goodness of the fit. Other models have either clearly worse verification scores or the goodness of the fit. Consequently, only the model C with the prior weight W_1 and the model G with the prior weight W_3 are included in the further analysis. The differences in the verification results are so small that it is not obvious which one of these two models is the best. The model G with prior weight W_3 represents stability in addition to wind direction and wind speed. However, the amount of data is so small that the real effect of stability may not improve the ELR model notably and does not give additional value to the model compared to the model C with only wind direction and wind speed as predictor variables.

5.3.3 Further investigation of the best models

The further analysis involve the sensitivity analysis for the *BSS*, the goodness of the fit and the P-values as well as the comparison of the reliability and the ROC diagrams.

The sensitivity for the *BS* is not calculated since the *BS* for different samples should not be compared.

The sensitivity analysis for the verification scores and the statistical testing results is performed with the bootstrap method without replacement. Slightly different subsets of the training data are used to create an ELR model a 100 times. These models are then verified with all the verification data. The training data includes 199 values on the whole and each ELR model is created with a random subset of 133 values that is approximately 2/3 of all the training data. A random noise is added to the original AROME/Harmonie wind speed of the subset. The purpose of the random noise is to vary the sample to create truly different training data of the existing data to simulate the situation for different years. We choose to use a random noise with the mean of 0 and the standard deviation of 0.5.

Other subset sizes and standard deviations for the random noise are examined as well. However, too small subset sizes, such as 60 or 100 values, causes P-values greater than 0.05. Also the verification scores and the goodness of the fit values are worse than for the subset of 133 values. Likewise, P-values are greater than 0.05 and both verification scores and the goodness of the fit are worse with standard deviations 1 and 2 compared to the standard deviation 0.5. Therefore, with less data and more deviation in the training sample, the ELR models get worse.

The sensitivity results show how much the verification results vary in different circumstances. The sensitivity scores analysed involve maximum values, minimum values, mean values and standard deviation of the verification scores and the goodness of the fit. The sensitivity scores for the 100 ELR models are presented in Tables 5.10 and 5.11.

Verification score	**Max**	**Min**	**Mean**	**Standard deviation**
BSS	0.818	0.788	0.806	0.00562
R^2	0.853	0.785	0.821	0.0142

Table 5.10: Sensitivity scores with the model C with the prior weight W_1.

The mean *BSS* of 0.806 in the sensitivity analysis (Table 5.10) is only slightly worse than the *BSS* of 0.811 for the initial ELR model (Table 5.9) and approximately the same quantity as for the other models in Table 5.9. Additionally, the maximum *BSS* of 0.818 is better than the *BSS* for any of the models in Table 5.9. However, the minimum *BSS* of 0.788 is notably worse. The mean goodness of the fit value of 0.821 in the sensitivity analysis is equally good as with the initial ELR model, being just

0.01 smaller. The maximum goodness of the fit value of 0.853 is much better than that for any of the models in Table 5.9. However, the minimum goodness of the fit value of 0.785 is much worse than 0.822.

All in all, the verification scores and the goodness of the fit values in the sensitivity analysis do not vary much compared to the initial ELR model C with the prior weight W_1 and the maximum values were even better than the scores for other models in Table 5.9 in spite of the smaller amount of data used in the sensitivity analysis. In addition, all the P-values are zero with two decimal accuracy so all model parameters are acceptable in the sensitivity analysis for the model C with prior weight W_1.

Verification score	**Max**	**Min**	**Mean**	**Standard deviation**
BSS	0.817	0.765	0.798	0.00971
R^2	0.868	0.777	0.833	0.0189

Table 5.11: Sensitivity scores with the model G with the prior weight W_3.

The mean *BSS* of 0.798 in the sensitivity analysis is clearly worse than the *BSS* of 0.809 for the initial ELR model (Table 5.9) and the minimum *BSS* of 0.765 is even worse. The maximum *BSS* of 0.817 is better than the *BSS* for any of the models in Table 5.9 but the mean *BSS* in the sensitivity analysis is worse than the *BSS* for the models in Table 5.9. The mean goodness of the fit value of 0.833 in the sensitivity analysis is slightly better than the goodness of the fit for the models in Table 5.9, the maximum goodness of the fit being even better. However, the minimum goodness of the fit of 0.777 is somewhat worse than 0.827 for the initial ELR model.

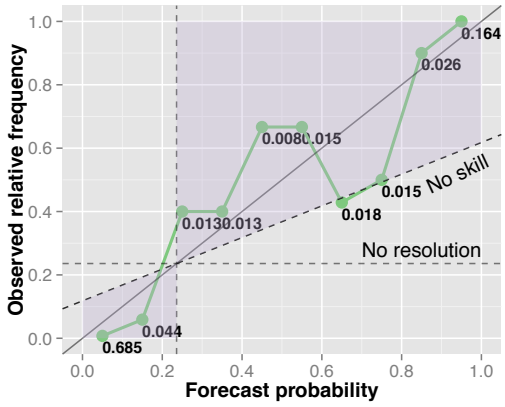
All in all, the *BSS* score is worse and the goodness of the fit is better in the sensitivity analysis compared to the initial ELR model G with the prior weight W_3. The mean goodness of the fit value in the sensitivity analysis is even better than for any of the models in Table 5.9 in spite of the smaller amount of data used in the sensitivity analysis but the mean *BSS* is worse than for any of the other models. Therefore, the model G with the prior weight W_3 is sensitive to variation. In addition, the P-values are zero with two decimal accuracy for other model parameters than stability. The stability variable is acceptable according to the mean and the minimum P-values but the maximum P-value 0.93 is bigger than 0.05. Therefore, stability is not an acceptable model parameter for all the 100 models in the sensitivity analysis, which is due to the small amount of data. Since the coefficient of the stability parameter is close to 0, it should not be involved in the model.

Comparing the sensitivity results for both of the models in the further analysis,

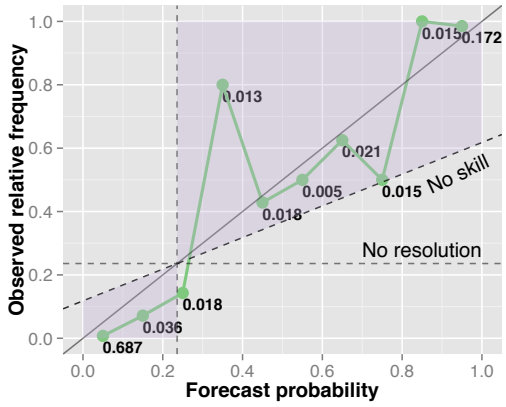
the *BSS* is better for the model C with the prior weight W_1 than for the model G with the prior weight W_3. The maximum *BSS* is equally good for both of the models so some of the 100 models are equally good as for the *BSS*. The maximum and the mean goodness of the fit are slightly better for the model G with the prior weight W_3 but the minimum goodness of the fit is slightly better for the model C with the prior weight W_1. The sensitivity results vary more for the model G with the prior weight W_3. In addition, the P-values are better for the model C with the prior weight W_1 than for the model G with the prior weight W_3. The sensitivity analysis therefore suggests that on average the model C with the prior weight W_1 works better than the model G with the prior weight W_3. However, we still analyse the reliability diagram and the relative operating characteristics that are calculated without the bootstrap method.

The reliability diagrams of the two models involved in the further analysis are illustrated in Figure 5.4. Instead of a sharpness diagram, the numbers next to the points show the relative amount of forecasts in the bin in question. Figure 5.4(a) shows that all the points are not close to the diagonal that stands for a perfectly reliable forecast. The first bin involves about 69% of the values so most of the probabilistic forecasts are below 0.1. The last bin stands for probabilistic forecasts above 0.9 and involves about 16% of the values. Other bins include notably fewer values, which make them more sensitive. The model partly over-forecasts and partly under-forecasts but most of the bins are close to the diagonal and the model clearly has resolution. The reliability for the model C with the prior weight W_1 is 0.00481. Figure 5.4(b) shows that all the points are not close to the diagonal either for this model. Likewise, the first bin involves about 69% of the values and the last bin includes about 17% of the values and other bins include notably fewer values, which make them more sensitive. The model mostly over-forecasts but the points are slightly closer to the diagonal on average comparing to the other model except the fourth point. This model also has resolution. The reliability for the model G with the prior weight W_3 is 0.00582.

It is not straight forward to analyse which reliability score is better. The reliability for the model C with the prior weight W_1 is 0.001 smaller than the reliability for the model G with the prior weight W_3. However, the bigger reliability for model G with prior weight W_3 is likely due to the fourth bin being far from the diagonal. Otherwise the reliability line for the model G with the prior weight W_3 is closer to the diagonal for the middle bins.



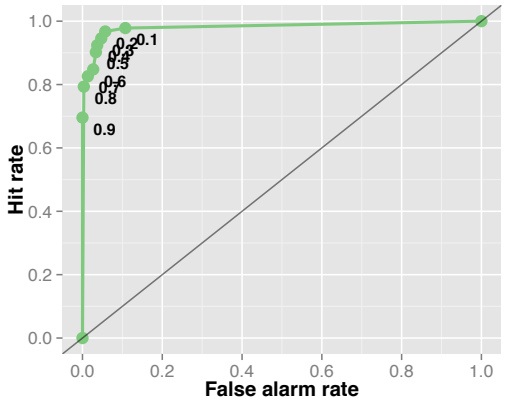
(a) Reliability plot for the model C with the prior weight W_1.



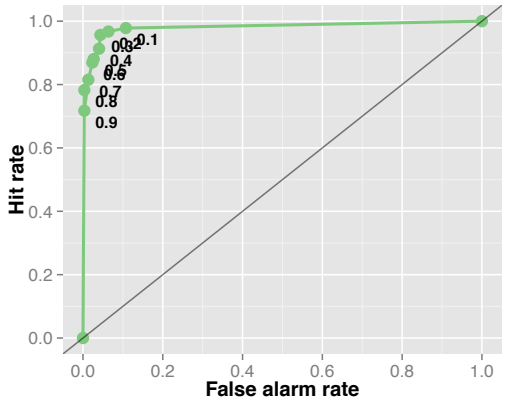
(b) Reliability plot for the model G with the prior weight W_3.

Figure 5.4: Reliability plots for the two models included in the further analysis.

The ROC area for both of the models is large, 0.990. Also, the ROC curves in Figure 5.5 look very similar to each other and the false alarm rates are small as expected considering the ROC values. Therefore, we cannot make any difference as for the ROC.



(a) The ROC plot for the model C with the prior weight W_1.



(b) The ROC plot for the model G with the prior weight W_3.

Figure 5.5: The ROC plots for the two models included in the further analysis.

To conclude, the model C with the prior weight W_1 is a better choice compared to the

model G with the prior weight W_3 based on the sensitivity results. There is no doubt that the model parameters would not be acceptable for the model C with the prior weight W_1. Reliability results are also better for the model C with the prior weight W_1 because the reliability score is smaller, only one bin has no skill and the model performs well for all the bins on average. Consequently, the best ELR model is the model C that involves the square root of the threshold, the AROME/Harmonie wind speed and the AROME/Harmonie wind direction category as predictor variables and the AROME/Harmonie wind speed as the prior weight.

5.3.4 Wind speed forecast and basic scores

Wind speed forecasts are extracted from the probabilities calculated with the chosen ELR model C with the prior weight W_1. Eventually, the basic scores for the error between the wind speed forecast and observations are analysed. The formula for the chosen ERL model presented with two decimal accuracy is

$$ln\left(\frac{p(q)}{1-p(q)}\right) = 6.79 - 5.20\sqrt{q} + 0.70x_1 + 1.27x_3,$$

where q is the threshold, x_1 is the AROME/Harmonie wind speed forecast and x_3 is the AROME/Harmonie wind direction category. Probability forecasts $p(q)$ are solved from the ELR equation for each threshold q. The produced probability distributions for each Julian day are then altered into wind speed forecast. In this thesis, the probability forecast means that the wind speed is above the given threshold q. To get the probability p that the wind speed is between two threshold values q_i and q_{i+1}, we calculate the difference

$$p = p_i - p_{i+1}.$$

The difference for the biggest threshold and infinity is simply the probability that the wind speed is above that threshold. To get more accurate results, we calculate the forecasts with threshold values between 0 and 14 at 0.1 intervals instead of using the thresholds that are used in the verification. As an example, the probability distribution for the Julian day 120, 30$^{\text{rd}}$ April, is shown in Figure 5.6. The height of the bars in Figure 5.6 represent the probability that the wind speed is between two threshold values. The curve is relatively smooth with the number of thresholds used.

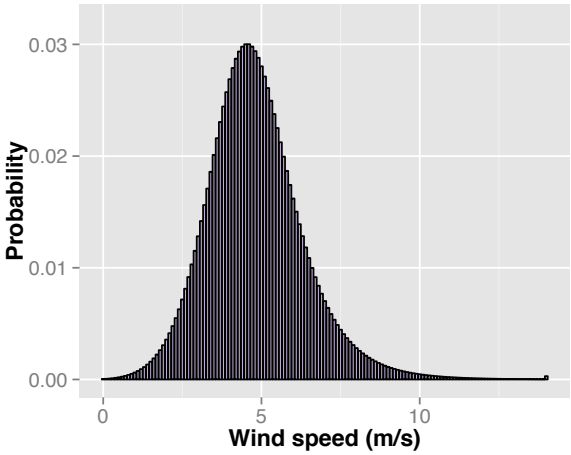


Figure 5.6: Probability distribution for wind speed on 30rd April forecasted with the model C with the prior weight W_1.

The most probable wind speed is the median wind speed of the probability distribution. The median wind speed is achieved by solving the threshold wind speed for which the probability is closest to 50%. The median wind speeds, in other words, the wind speed forecasts achieved with the chosen ELR model for the whole verification period are shown in Figure 5.7. The initial AROME/Harmonie forecasts and observations are included in the figure as well. The ELR forecast accuracy is one decimal due to the choice of the threshold interval.

Figure 5.7 shows that the ELR forecasts follow the observations better than the initial AROME/Harmonie forecasts. The ELR forecasts range between 2.8 – 9.7m/s. The AROME/Harmonie forecasts range between 2.3 – 13.3m/s and the observations range between 1.7m/s and 10.5m/s. The ELR model smooths the pikes of the initial AROME/Harmonie forecast staying slightly under the observation curve in case of high wind speeds. A systematic error is likely to occur in the Arome/Harmonie forecast because the AROME/Harmonie forecast curve is above the observation curve most of the time. The ERL model reduces this error. The basic scores — mean error, root mean square error, mean absolute error, absolute maximum error and absolute minimum error — for error between the ELR forecast and the observations and the AROME/Harmonie forecasts and the observations respectively are presented in Table 5.12. The accuracy is three significant values for other scores than the absolute maximum and minimum errors that are not more accurate than two significant values.

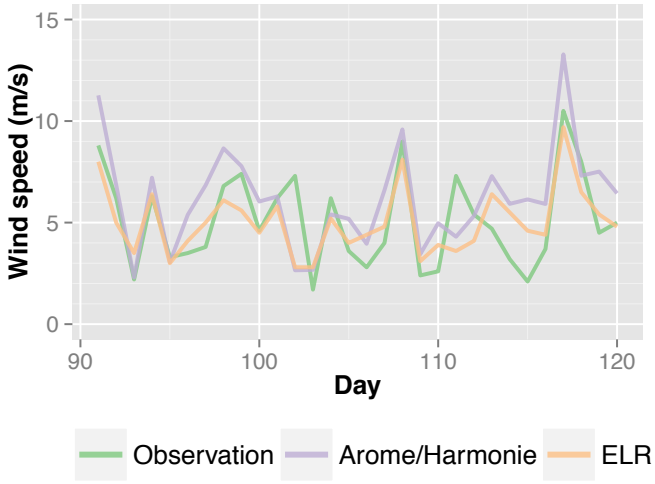


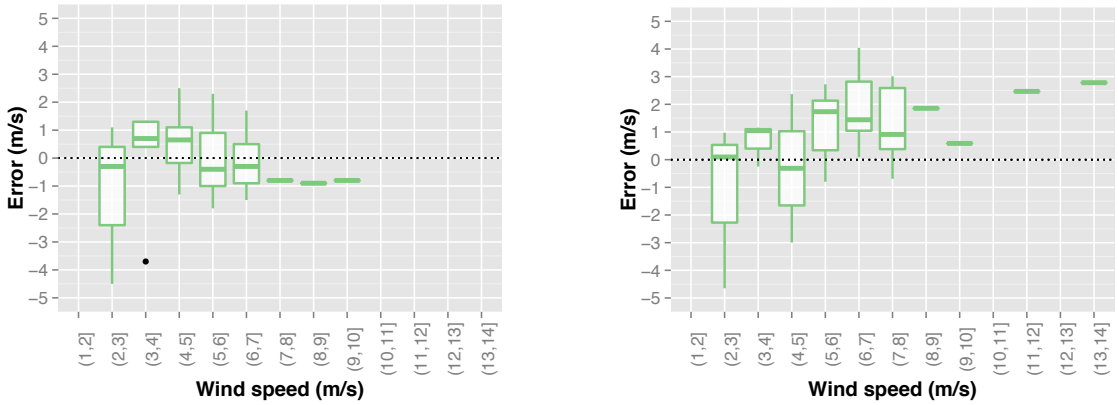
Figure 5.7: Wind speed for the verification period. The observed wind speed is illustrated with a green line, the initial AROME/Harmonie forecast with a violet line and the forecasted wind speed with the ELR model C with prior weight W_1 with a yellow line.

The basic scores for the error presented in Table 5.12 are smaller for the ELR model than for initial AROME/Harmonie forecast. The absolute maximum error of 4.5m/s for the ELR is only 0.01 smaller than for the AROME/Harmonie forecast. It occurs for the verification day 102 when the ELR model forecasts the wind speed to reduce suddenly one day too early. However, the *MAE* for ELR is 0.5m/s smaller than that for the AROME/Harmonie forecast, which means 29% smaller mean absolute error. The *RMSE* is also 0.5m/s smaller, which is 24% smaller root mean square error, which is a remarkable improvement to the forecasts. Therefore, the basic suggest that the ELR model created for the Olkiluoto mast produces better wind speed forecasts than the AROME/Harmonie forecasts.

Model	**ME**	**RMSE**	**MAE**	**Max**	**Min**
ELR	-0.0667	1.58	1.21	4.5	0.10
AROME/Harmonie	1.08	2.11	1.71	4.6	0.08

Table 5.12: Basic scores for the ELR and the AROME/Harmonie forecasts.

The forecast errors in quantiles for the ELR model and the AROME/Harmonie model (Figure 5.8) support the results of the basic scores of the ELR model to reduce the error of the AROME/Harmonie forecast. The bars are shorter and closer to zero for the ELR model since they are only at about 1m/s range from zero whereas the error bars for the AROME/Harmonie forecast are at about 2m/s range from zero. The biggest errors for the ELR model are for wind speeds around 3m/s whereas the biggest errors for the AROME/Harmonie model are for wind speeds 3m/s and 5m/s. The errors for wind speeds greater than 8m/s do not range much for the ELR model and the errors for wind speeds greater than 9m/s do not range much for AROME/Harmonie forecasts.



(a) Errors between the ELR forecasts and the observations in quantiles.

(b) Errors between the AROME/Harmonie forecasts and the observations in quantiles.

Figure 5.8: Forecast errors in quantiles for the verification period.

5.4 Regression estimation of event probabilities model

The regression estimation of event probabilities (REEP) model is created with the same training and prediction data sets as the ELR model. The threshold values used for the REEP model are $q = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13$ and 14. The binary values for the observations are calculated for these thresholds. The REEP equation (Equation 4.8) involve only one predictor variable x that is the AROME/Harmonie wind speed. The *glm* function of R is used to calculate the intercept and the coefficient

for the predictor variable by using the training data. In this procedure, the binary observation values are matched to the AROME/Harmonie wind speeds directly without a logit link function. The REEP model made for the training data gets a form for each threshold separately and is presented with two decimal accuracy as follows

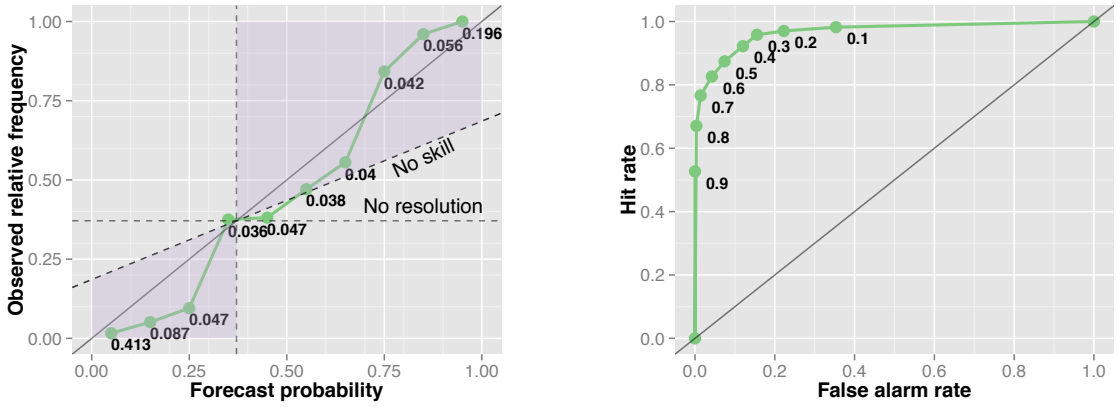
$$
\begin{aligned}
q = 0 &: y_0 = 1.00 + 0.00x, \\
q = 1 &: y_1 = 0.93 + 0.01x, \\
q = 2 &: y_2 = 0.75 + 0.03x, \\
q = 3 &: y_3 = 0.42 + 0.06x, \\
q = 4 &: y_4 = 0.17 + 0.08x, \\
q = 5 &: y_5 = -0.14 + 0.11x, \\
q = 6 &: y_6 = -0.26 + 0.10x, \\
q = 7 &: y_7 = -0.33 + 0.09x, \\
q = 8 &: y_8 = -0.29 + 0.06x, \\
q = 9 &: y_9 = -0.27 + 0.05x, \\
q = 10 &: y_{10} = -0.22 + 0.04x, \\
q = 11 &: y_{11} = -0.16 + 0.03x, \\
q = 12 &: y_{12} = -0.13 + 0.03x, \\
q = 13 &: y_{13} = -0.10 + 0.02x, \\
q = 14 &: y_{14} = -0.03 + 0.00x.
\end{aligned}
\tag{5.1}
$$

The probabilistic REEP forecasts for the prediction data are solved with either *predict* function or directly from the REEP functions for each threshold using the correspondent coefficients a_i and b_i. The probability values below 0 are altered to 0 and the values above 1 are altered to 1. The verification scores are calculated with the *verify* function of R to verify the new forecasts against the binary observations of the prediction data. The verification scores are shown in Table 5.13. There is no reason to calculate the goodness of the fit for all the REEP equations because several goodness of the fit cannot be compared simply with the goodness of the fit of the ELR model. Instead, the reliability values abbreviated with Rel and the ROC area values are added to the table.

Model	BS	BSS	Rel	ROC area
REEP	0.0660	0.717	0.00480	0.967
ELR	0.0341	0.811	0.00481	0.990

Table 5.13: Verification scores for the REEP and the ELR model.

Table 5.13 shows that the *BSS* and the *BS* for the REEP model are clearly worse than the *BSS* and the *BS* for the ELR model. The reliability score is equally good for both models with four decimal accuracy. The ROC value for the REEP model is relatively good compared to that of the ELR model — it is only 0.023 smaller. Consequently, according to the verification scores, the ELR models is better, although it must be taken into consideration that the REEP model is much simpler and it does not cover all possible thresholds. The reliability and ROC diagrams for the REEP model are presented in Figure 5.9.



(a) Reliability plot for the REEP model.

(b) ROC plot for the REEP model.

Figure 5.9: Reliability and ROC for the REEP model.

Figure 5.9(a) shows that almost all the forecast probability bin points are relatively close to the diagonal. The model mostly over-forecasts but it still has resolution. The first bin close to zero includes about 41% of the values and the last bin includes about 20% of the values. Other bins include notably fewer values, which make them more sensitive. The ROC diagram is good for the REEP model as expected due to the large ROC area.

The probability forecasts are calculated for the REEP model only with the threshold values between 0 and 14 at 1 intervals. As an example, the probability distribution

for the Julian day 120, the 30rd of April, is shown in Figure 5.10. The probability distribution curve is more robust for the REEP model than the probability distribution curve for the ELR model at least partly due to the smaller amount of thresholds.

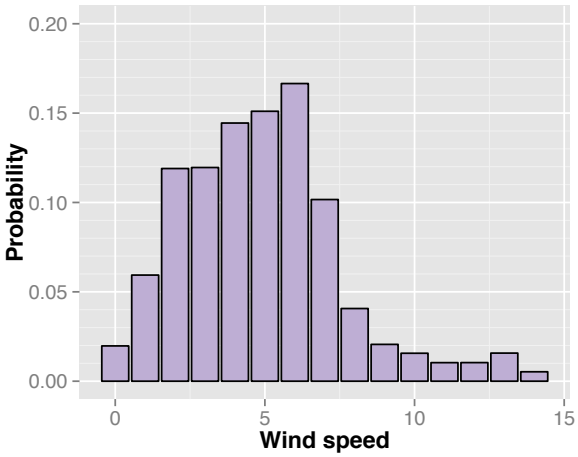


Figure 5.10: Probability distribution for wind speed on 30rd April forecasted with the REEP model.

Probability forecasts are altered into wind speed values the same way as for the ELR model. The most probable wind speed is the median wind speed of the probability distribution. The wind speed forecasts created with the REEP model for the whole verification period are shown and compared to the initial AROME/Harmonie forecast and the observations in Figure 5.11.

Figure 5.11 shows that the REEP range very little, only between 2.0 – 7.0 m/s, while the observations range between 1.7 m/s and 10.5 m/s. So, the REEP forecasts are quite good for very average wind speeds but not for very low or high wind speeds at least with this little data. The AROME/Harmonie forecasts follow the observations better, especially at higher wind speeds. The basic scores for the error between the REEP forecasts and the observations and the AROME/Harmonie forecasts and the observations respectively are presented in Table 5.14.

Table 5.14 presents that the *RMSE* and the *MAE* are only slightly smaller with the REEP model than with the initial AROME/Harmonie forecast. The absolute maximum error and the *ME* is smaller for the REEP model likely due to averaging of the model. Regardless that the basic scores for the AROME/Harmonie forecasts are slightly worse, the AROME/Harmonie forecasts follow the observation variations

much better.

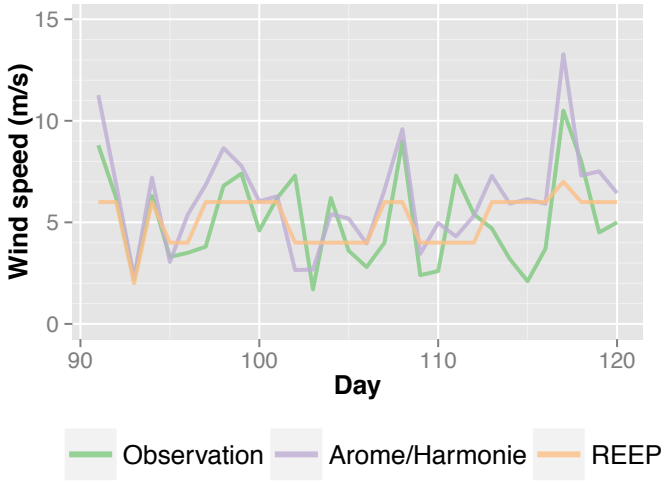
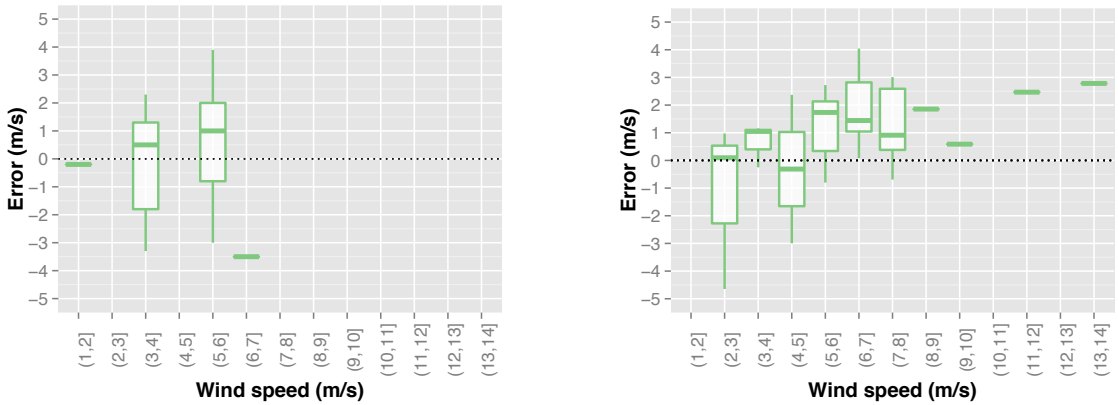


Figure 5.11: Wind speed for the verification period. The observed wind speed is with green, the initial AROME/Harmonie forecast with violet and the REEP forecast with yellow.

Model	ME	RMSE	MAE	Max	Min
REEP	0.0633	2.04	1.70	3.9	0.20
AROME/Harmonie	1.08	2.11	1.71	4.6	0.08

Table 5.14: Basic scores for the REEP and the AROME/Harmonie.

The forecast errors in quantiles for the REEP model and the AROME/Harmonie model for the verification period are shown in Figure 5.12. Considering that the REEP forecasts range much less than the initial AROME/Harmonie forecasts, the errors are about of the same quantity. The error for wind speeds at 7m/s is large and negative for the REEP model whereas that for the AROME/Harmonie model is positive and smaller. The REEP model has only one forecast at 7m/s due to the pike in both observed wind speed and AROME/Harmonie forecast. Therefore, the REEP model does not reduce the error of the AROME/Harmonie forecast, only averages the initial AROME/Harmonie forecast.



(a) Errors between the REEP forecasts and the observation in quantiles.

(b) Errors between the AROME/Harmonie forecasts and the observations in quantiles.

Figure 5.12: Forecast error in quantiles for the verification period.

5.5 Linear model

In the deterministic linear model (LM) the observation values are not altered to probabilistic binary values. Consequently, there are no threshold values either. The same training and prediction data are used as for the ELR and the REEP models. The LM is formed with the *glm* function of R which calculates the intercept and the coefficient for the predictor variable. The observed wind speed values are matched to the AROME/Harmonie wind speed values directly without a logit link function. The LM created with the training data gets a form

$$y = 0.94 + 0.70x.$$

The achieved forecast values produced with the LM are directly wind speed values so no verification scores can be computed to compare with the ELR and the REEP models. There is no reason to compare the goodness of the fit because it was not calculated for the REEP model either. The wind speed forecasts created with the LM are shown and compared to the initial AROME/Harmonie forecasts and the observations in Figure 5.13.

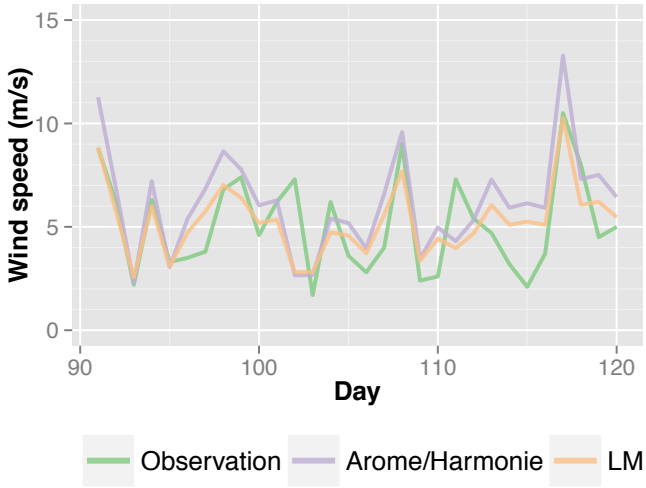


Figure 5.13: Wind speed for the verification period. The observed wind speed is with green, the initial AROME/Harmonie forecast with violet and the LM forecast with yellow.

Figure 5.13 shows that the LM forecasts follow the observations really well. The LM forecast ranges between 2.6 – 10.3 m/s while the observations range between 1.7 m/s and 10.5 m/s. The LM forecast curve stays under the AROME/Harmonie forecast curve all the time, which suggest that the LM reduces the systematic error of the AROME/Harmonie forecast. The basic scores for error between the LM forecasts and the observations and the AROME/Harmonie forecasts and the observations respectively are presented in Table 5.15.

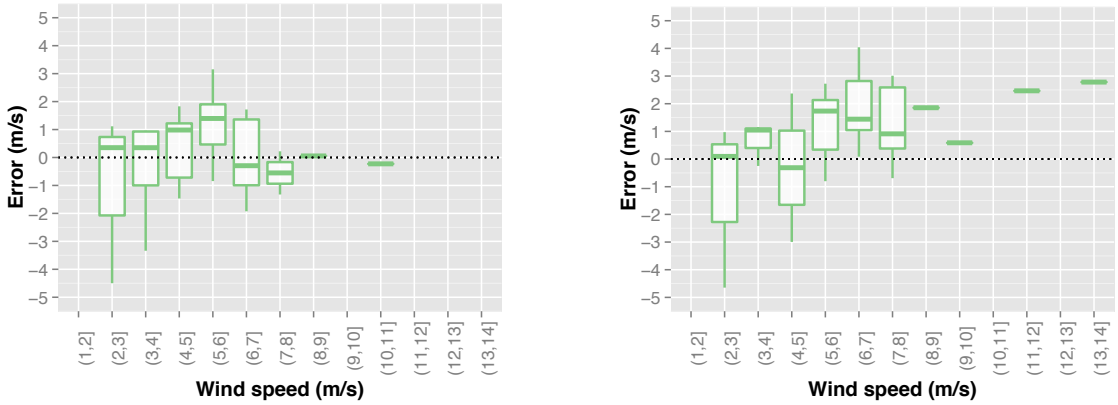
Model	ME	RMSE	MAE	Max	Min
LM	0.182	1.64	1.27	4.5	0.06
AROME/Harmonie	1.08	2.11	1.71	4.6	0.08

Table 5.15: Basic scores for the LM and the AROME/Harmonie.

Table 5.15 presents that all the basic scores are smaller for the LM than for the initial AROME/Harmonie forecasts. Both the basic scores and the graphical analysis suggest that the LM reduces the error of the AROME/Harmonie forecasts.

The forecast errors in quantiles for the LM model and the AROME/Harmonie model for the verification period shown in Figure 5.14. Figure 5.14 supports the

results of the basic scores that the LM reduces the error of the AROME/Harmonie forecast.



(a) Error between LM forecast and observation in quantiles.

(b) Error between AROME/Harmonie forecast and observation in quantiles.

Figure 5.14: Forecast error in quantiles.

5.6 Comparison of models

The chosen ELR model is compared to the reference models as well as to the initial AROME/Harmonie forecasts by evaluating the size of the errors with the calculated basic scores and the graphical analysis. Also, the verification scores for the ELR model and the REEP model are taken into consideration. The basic scores for the errors are gathered in Table 5.16.

Model	ME	RMSE	MAE	Max	Min
ELR	-0.0667	1.58	1.21	4.5	0.10
REEP	0.0633	2.04	1.70	3.9	0.20
LM	0.182	1.64	1.27	4.5	0.056
AROME/Harmonie	1.08	2.11	1.71	4.6	0.075

Table 5.16: The basic scores for the ELR, the REEP, the LM and the AROME/Harmonie forecasts.

Table 5.16 shows that the *MAE* and the *RMSE* are smallest for the ELR model. All the basic scores are notably better for the ELR model especially compared to the

initial AROME/Harmonie forecast. The absolute mean error is 29% and the root mean square error is 24% smaller for the ELR forecast compared to the AROME/Harmonie forecasts. It is advantageous to have a small *RMSE* because large errors have a remarkable effect on the yield of the wind power plant. The ELR model also smooths the pikes of the AROME/Harmonie forecast (Figure 5.7), which is essential since the electricity production must be stopped with wind speeds greater than 25m/s. The ELR forecast curve and the observation curve are well compatible. Both the basic scores and the graphical evaluation suggest that the ELR model reduces the systematic error of the AROME/Harmonie forecast. Additionally, the ELR model reduces the error for all quantiles and the model works approximately equally well for each quantile (Figure 5.8). The ELR model was supposed to forecast all wind speeds equally well so this result supports the use of the ELR model.

The LM has the second best basic scores (Table 5.16), almost equally good as for the ELR model. The LM forecasts fit the observation curve slightly better than the AROME/Harmonie forecasts because the LM forecasts smooth the pikes (Figure 5.13). Also the LM reduces the systematic error of the AROME/Harmonie forecast. Considering the small amount of data, the differences between the LM and the ELR model are not very big. However, the LM model is purely linear whereas the ELR model is nonlinear. Consequently, the ELR model improves the results by taking into account the nonlinearity of the error. The ELR model has also other parameters than wind speed that improve the predictability.

Despite the small *ME* and the small absolute maximum error, the *RMSE* and the *MAE* are only slightly smaller with the REEP model than with the AROME/Harmonie forecast. The REEP forecast curve does not follow the observation curve well (Figure 5.11) and therefore the REEP model gives little value in forecasting wind speeds. In addition, verification results (Table 5.13) show that the *BSS* and the *BS* for the REEP model are clearly worse than those for the ELR model. The REEP model could possibly perform better with several threshold values but it would be laborious to calculate an equation for each threshold.

Considering both the basic scores and the graphical analysis, the ELR model is clearly better than the initial AROME/Harmonie forecast or the reference models. The REEP model averages the AROME/Harmonie forecast and has little value. The LM has both better basic scores than the REEP model and it is simpler. Thus, it is not reasonable to use the REEP model for forecasting wind speeds. Regardless that the LM rises above the REEP model, it has slightly worse basic scores than the ELR model.

Chapter 6

Evaluation of the model

The chosen ELR model with the square root of the threshold wind speed, the initial AROME/Harmonie wind speed and the AROME/Harmonie wind direction category as predictor variables outperforms the reference models and the initial AROME/Harmonie forecasts. All the basic scores for the error are clearly smallest for the ELR forecasts. The absolute mean error is 29% smaller and the root mean square error is 24% smaller for the ELR forecasts compared to the AROME/Harmonie forecasts. The graphical evaluation also suggests that the ELR forecasts reduce the systematic positive error of the AROME/Harmonie forecast (Figure 5.7). The ELR model reduces the error for all quantiles and the model works approximately equally well for each quantile (Figure 5.8). Consequently, the ELR model gives more accurate results with less uncertainty. Nevertheless, the ELR model could be even better, so the weaknesses of the model and the aspects for future development of the model are analysed.

First of all, the ELR forecast curve is highly dependent on the shape of the initial AROME/Harmonie forecast curve. If the AROME/Harmonie forecast curve does not fit the observation curve, the ELR forecast curve fits neither, despite the fact that the ELR model reduces the systematic error. The numerical AROME/Harmonie forecast is only an average forecast with forecast errors. The forecast error in the AROME/Harmonie forecast results from an imperfect initial state regardless data assimilation, and the imperfect numerical model. The initial forecast error affects the results of the ELR model. Resolution improvements and better initial data obtained with new observation methods, such as new satellites and weather radars, could reduce these forecast errors of the AROME/Harmonie model. Initial errors are likely to exist in the observation data as well, which affect the verification results of the ELR model. It would be worth considering to reduce the bias from the AROME/Harmonie forecasts before using them for creating the ELR model.

The most significant problem with the modelling is the limited amount of data available. The data covers only 229 days from January to the middle of August, which is less than a year. It would be reasonable to have data from at least three years because wind changes in long-term periods. In an optimal situation, we would have at least two years' data for the training and one year's data for the verification. Thus, the choice of the training and verification data also affects the results although the verification period chosen represents the average conditions for the data. With more data, the error scores for the model would supposedly be even smaller, seasonal variation could be taken into account and the data would include more high wind speeds. In this thesis, the data involves only few high wind speeds, which makes the results sensitive. According to the literature review, the ELR method works better for extreme values, so the ELR would have potential for modelling high wind speeds.

The limited amount of data affects also model choices because the results are sensitive to small changes, such as involving single data values into the training data or different categories. The sensitivity analysis made for the two ELR models point out that the verification scores vary substantially with slightly different data sets. The choice of the ELR model could be blamed for comparing too accurate results and the verification results and statistical results could be prioritised differently. However, Table 5.9 shows that the *BSS* scores are at least 0.803 for all the best models and the goodness of the fit is at least 0.806. The changes are not big and all the goodness of the fit scores are relatively high, which implies that the results have skill. The verification scores and the statistical results are considered together, which is a reasonable way to compare the models.

Another problem with the AROME/Harmonie data is that data from the ground level is not available. Besides, the vertical resolution of the AROME/Harmonie could be better in order to provide data at more levels. Especially the stability category variable could have more impact if it would be possible to take an actual change around the chosen AROME/Harmonie model height. In this thesis, the stability variable is created with values at the model height (30m) and at the next height above (47m) instead. Moreover, the AROME/Harmonie model heights do not match with those of the observations because there is a systematic error in the AROME/Harmonie forecasts.

The chosen ELR model parameters appear to be reasonable and physically justified. The AROME/Harmonie wind speed forecast is naturally the principal parameter. The categorical wind speed variable improves the verification and the statistical testing results with all the prior weights. The improvement is expected because the wind mast is situated on the coastal area so wind speed is different from off-side and on-side

directions. The boundary values for wind speed categories also cohere with the map (Figure 3.4). The stability category variable has some positive effect on the results but less than the wind direction category variable. This is likely due to the night-time data because stability is smaller during nigh-time and bigger during daytime. Therefore, it is evident that the stability parameter does not improve the results much because there are only few unstable situations in the data. However, stability can have value when forecasting daytime wind speeds. The temperature and temperature difference parameters are acceptable and improve the results for two prior weights. Nevertheless, the best models do not include temperature difference and temperature only together with wind direction. The AROME/Harmonie forecasts already include physical parameters, such as temperature, which is likely a reason that temperature does not have a notable effect on the results. According to the literary review, the square root of threshold improves the results for forecasting precipitation. It improves the results for forecasting wind speed as well compared to having a linear function of thresholds.

More alternative models could be created by taking into account more combinations of model parameters and functions of them. Also totally new parameters could be added if more AROME/Harmonie forecasts were available. Regardless, both the choice of the function of the threshold *g(c)* and the model parameters are totally empirical. There could exist even better models, but it would take plenty of time to try countless alternatives. In addition to model parameters, the choice of the prior weights also affects the results. The use of wind speed and functions of it as prior weights is physically understandable but it could be possible to find other prior weights that would work even better. With this much data, it is reasonable to build a simplified model with not too many parameters because the small improvements with a more complicated model could not be justified.

To conclude, the errors in the ELR model are due to the errors in the initial observations and the AROME/Harmonie forecasts, the small amount of data, the choice of parameters not being optimal and the model not being perfect. It is likely to get different results with different or bigger amount of data but the improvements to the AROME/Harmonie forecasts do not seem to be random with this amount of data either. The AROME/Harmonie forecasts already take into account physical aspects and the ERL model strengthens the physical dependencies. All in all, the ELR model created is not a perfect model but in spite of its deficiencies it improves the results, which is the most important goal to achieve in this thesis. The ELR model can also be updated when getting more data. Eventually, we conclude that the ELR method works for wind speed forecasting and improves the results for the Olkiluoto mast.

Chapter 7

Conclusions

The ELR model that gives the best results includes the square root of the threshold wind speed, the AROME/Harmonie wind speed and the AROME/Harmonie wind direction category as model parameters, the prior weight being the AROME/Harmonie wind speed. The thorough investigation of the verification scores and the statistical results shows that this model parameter combination and prior weight outperform the other parameter combinations and prior weights examined. The model with the stability category as an additional parameter and the AROME/Harmonie wind speed squared as the prior weight works nearly equally well. Nevertheless, that model is more sensitive since there are only few values in the category for unstable atmosphere.

The differences between the alternative ELR models would become clearer with a longer period of data. It is essential to gather more data in the future studies, preferably during a period of three years. In spite of the small amount of data, it is shown that the new forecasts are more accurate than the initial AROME/Harmonie forecasts. The wind speed forecasts calculated from the probabilistic ELR forecasts have smaller error scores compared to the initial AROME/Harmonie wind speed forecasts and the wind speed forecasts of the reference models. Moreover, the graphical evaluation suggests that the ELR forecasts follow well the observations and smooth the pikes of the AROME/Harmonie forecasts for high wind speed values. Regardless the deficiencies of the AROME/Harmonie model being reflected to the ELR model, the ELR model reduces the error and therefore the uncertainty with 29% smaller mean absolute error and 24% smaller root mean square error.

The ELR model created in this thesis is simple, built with little data and limited to only one location and forecasts only at 03 UTC. The model needs to be extended to cover all hours of a day before the results can be launched for commercial use. There could also be a moderately different model for day and night because of different

stability conditions. In addition, models could be formed for various heights and locations with observation data available. The coefficients for parameters and the wind directions categories vary according to location. Nonetheless, the models for different wind masts are likely very similar to the model created in this thesis. In the future, the model can probably be extended to cover all Finland.

The wind speed forecasts generated with the ELR model can be utilised to create more accurate wind power forecasts. The new forecasts will substantially reduce the uncertainty of the electricity production although the actual power of a wind mast must be calculated separately with the help of power curves, especially in a wind power park with blind spots. The more wind power is inserted in the grid, the more important it is to forecast wind power accurately with different time horizons both from the producer and the system operator point of view. Better forecasts decrease the need for balancing power and thus the balancing costs as well. Additionally, the information that the ELR model does not over-forecast high wind speeds is essential regarding the need to stop the wind power production in case of storms.

Forecasting electricity production for the Elspot market in Nord Pool Spot requires longer than 12 hour forecasts. Several new models should therefore be calculated with longer AROME/Harmonie forecast lengths. In the best case, the ELR model also works for longer AROME/Harmonie forecasts. The 3 hour forecasts made could theoretically be used for the Elbas market but it is not yet profitable in Finland. Nevertheless, the share of wind power will increase in Finland due to politics supporting renewable energies. It is estimated to be 6TWh in 2020 whereupon it would be cost-effective to sell wind power in the Elbas market.

The results show evidence of the ELR method being an opportunity to make more accurate wind speed forecasts for power production. The major finding that the ELR method works for forecasting wind speed by reducing uncertainty significantly is also crucial for research and applications for forecasting wind speeds. The ELR model created in this thesis improves the short-term wind speed forecasts for the Olkiluoto mast and similar models can likely be formed for different locations, heights and time resolutions. The extension of the ELR model and the implementation of applications require more data and further analysis. However, the ELR model created in this thesis is a good basis for further research.

Bibliography

Busby, R. L. (2012). *Wind Power - The Industry Grows Up.* PennWell, Oklahoma, USA, 1 edition.

Coiffier, J. (2011). *Fundamentals of Numerical Weather Prediction.* Cambridge University Press, New York, USA, 1 edition.

Finnish Energy Industries (3.3.2014). Electricity Generation. `http://energia.fi/energia-ja-ymparisto/sahkontuotanto`.

Galperin, B., Sukoriansky, S., and Anderson, P. S. (2007). On the Critical Richarson Number in Stably Stratified Turbulence. *Atmospheric Science Letters*, 8:65–69.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer Sciene+Business Media, New York, USA, 2 edition.

HIRLAM (13.2.2013). HIRLAM. `http://hirlam.org`.

Holton, J. R. (1992). *An Introduction to Dynamic Meteorology.* Academic Press, San Diego, USA, 3 edition.

Holttinen, H. (2004). The Impact of Large Scale Wind Power Production on the Nordic Electricity system. *VTT Publications 554, Doctoral Dissertation, Helsinki University of Technology.*

Holttinen, H. and Koreneff, G. (2012). Imbalance Costs of Wind Power for a Hydra Power Producer in Finland. *Wind Engineering*, 36:53–68.

Holttinen, H., Meibom, P., Orths, A., Lange, B., O'Malley, M., Tande, J., Estanqueiro, A., Gomez, E., Söder, L., Strbac, G., Smith, J., and van Hulle, F. (2011). Impacts of Large Amounts of Wind Power on Design and Operation of Power Systems, Results of IEA Collaboration. *Wind Energy*, 14:179–192.

Holttinen, H., Miettinen, J., and Sillanpää, S. (2013). Wind Power Forecasting Accuracy and Uncertainty in Finland. *VTT Technology 95*.

Jolliffe, I. T. and Stephenson, D. B., editors (2003). *Forecast Verification, A Practitioner's Guide in Atmospheric Science*. Wiley, West Sussex, UK, 1 edition.

Kaimal, J. C. and Finnigan, J. J. (1994). *Atmospheric Boundary Layer Flows: Their Structure and Measurement*. Oxford University Press, New York, USA, 1 edition.

Laininen, P. (2000). *Tilastollisen analyysin perusteet*. University Press Finland Ltd., Helsinki, Finland, 3 edition.

Manwell, J. F., McGowan, J. G., and Rogers, A. L. (2009). *Wind Energy Explained*. John Wiley and Sons Ltd., Chichester, UK, 2 edition.

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S. (2013). Heteroscedastic Extended Logistic Regression for Post-Processing of Ensemble Guidance. *Monthly Weather Review*, 142.

National Land Survey of Finland (18.8.2014). National Land Survey of Finland. `http://kansalaisen.karttapaikka.fi/`.

Nord Pool Spot (4.3.2014). Nord Pool Spot. `http://www.nordpoolspot.com`.

Nurmi, P. (14.1.2014). Forecast Verification. `http://www.eumetcal.org/resources/ukmeteocal/verification/www/english/courses/msgcrs/index.htm`.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V. (2010). The AROME-France Convective-Scale Operational Model. *Monthly Weather Review*, 139:976–991.

The Centre for Australian Weather and Climate Research (26.11.2013). Forecast Verification: Issues, Methods and FAQ. `http://www.cawcr.gov.au/projects/verification/`.

The Finnish Meteorological Institute (25.2.2014a). Sääennustedata. `http://ilmatieteenlaitos.fi/avoin-data-saaennustedata-hirlam`.

The Finnish Meteorological Institute (27.8.2014b). The Finnish Meteorological Institute. `http://en.ilmatieteenlaitos.fi`.

The Finnish Wind Power Association (4.2.2014). The Finnish Wind Power Association. `http://www.tuulivoimayhdistys.fi/`.

The Ministry of Employment and the Economy (Finland) (18.12.2013b). The Finnish Wind Atlas. `http://www.tuuliatlas.fi/fi/index.html`.

The Ministry of Employment and the Economy (Finland) (2013a). National Energy and Climat Strategy. *MEE Publications*.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*. Academic Press, INC., London, UK, 2 edition.

Wilks, D. S. (2009). Extending Logistic Regression to Provide Full-Probability-Distribution MOS Forecasts. *Meteorological Applications*, 16:361–368.

Wilks, D. S. and Hamill, T. M. (2007). Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Reviews*, 135:2379–2390.

Appendix A

Wind direction categories

Onshore [°]	**N**	**Offshore** [°]	**N**
0 – 230	161	230 – 360	68
0 – 240	166	240 – 360	63
0 – 250	170	250 – 360	59
10 – 230	157	230 – 360, 0 – 10	72
10 – 240	162	240 – 360, 0 – 10	67
10 – 250	166	250 – 360, 0 – 10	63

Table A.1: The number of wind direction values N in each category.

Onshore [°]	**Offshore** [°]	**BS**	**BSS**	$\mathbf{R^2}$	**P-value for** x_3
0 – 230	230 – 360	0.0357	0.802	0.819	0.04
0 – 240	240 – 360	0.0353	0.804	0.819	0.05
0 – 250	250 – 360	0.0354	0.804	0.819	0.05
10 – 230	230 – 360, 0 – 10	0.0351	0.805	0.819	0.05
10 – 240	240 – 360, 0 – 10	0.0346	0.808	0.819	0.07
10 – 250	250 – 360, 0 – 10	0.0353	0.804	0.818	0.07

Table A.2: Verification scores and the goodness of the fit for different wind category values with prior weight W_0.

Onshore [°]	**Offshore** [°]	**BS**	**BSS**	$\mathbf{R^2}$	**P-value for** x_3
0 – 230	230 – 360	0.0356	0.802	0.823	0
0 – 240	240 – 360	0.0341	0.811	0.822	0
0 – 250	250 – 360	0.0339	0.812	0.821	0
10 – 230	230 – 360, 0 – 10	0.0348	0.807	0.822	0
10 – 240	240 – 360, 0 – 10	0.0340	0.811	0.821	0
10 – 250	250 – 360, 0 – 10	0.0338	0.812	0.820	0

Table A.3: Verification scores and the goodness of the fit for different wind category values with prior weight W_1.

Onshore [°]	**Offshore** [°]	**BS**	**BSS**	$\mathbf{R^2}$	**P-value for** x_3
0 – 230	230 – 360	0.0348	0.807	0.818	0.00
0 – 240	240 – 360	0.0344	0.809	0.818	0.01
0 – 250	250 – 360	0.0350	0.806	0.817	0.01
10 – 230	230 – 360, 0 – 10	0.0346	0.808	0.818	0.01
10 – 240	240 – 360, 0 – 10	0.0347	0.807	0.817	0.01
10 – 250	250 – 360, 0 – 10	0.0348	0.807	0.817	0.02

Table A.4: Verification scores and the goodness of the fit for different wind category values with prior weight W_2.

Onshore [°]	**Offshore** [°]	**BS**	**BSS**	$\mathbf{R^2}$	**P-value for** x_3
0 – 230	230 – 360	0.0357	0.802	0.829	0.00
0 – 240	240 – 360	0.0348	0.807	0.826	0.00
0 – 250	250 – 360	0.0358	0.801	0.825	0.00
10 – 230	230 – 360, 0 – 10	0.0360	0.800	0.827	0.00
10 – 240	240 – 360, 0 – 10	0.0341	0.811	0.824	0.00
10 – 250	250 – 360, 0 – 10	0.0350	0.806	0.823	0.00

Table A.5: Verification scores and the goodness of the fit for different wind category values with prior weight W_3.

Onshore [°]	**Offshore** [°]	**BS**	**BSS**	$\mathbf{R^2}$	**P-value for** x_3
0 – 230	230 – 360	0.0376	0.792	0.834	0.00
0 – 240	240 – 360	0.0355	0.803	0.830	0.00
0 – 250	250 – 360	0.0355	0.803	0.828	0.00
10 – 230	230 – 360, 0 – 10	0.0374	0.792	0.832	0.00
10 – 240	240 – 360, 0 – 10	0.0357	0.802	0.828	0.00
10 – 250	250 – 360, 0 – 10	0.0353	0.804	0.826	0.00

Table A.6: Verification scores for different wind category values with prior weight W_4.